A Random Variables and Probability Distributions

A.1 Distribution Functions and Expectation

The distribution function F of a random variable X is defined by

$$F(x) = P[X \le x] \tag{A.1.1}$$

for all real x. The following properties are direct consequences of (A.1.1):

1. F is nondecreasing, i.e., $F(x) \le F(y)$ if $x \le y$.
2. F is right continuous, i.e., $F(y) \downarrow F(x)$ as $y \downarrow x$.
3. $F(x) \to 1$ and $F(y) \to 0$ as $x \to \infty$ and $y \to -\infty$, respectively.

Conversely, any function that satisfies properties 1–3 is the distribution function of some random variable.

Most of the commonly encountered distribution functions F can be expressed either as

$$F(x) = \int_{-\infty}^{x} f(y)dy \tag{A.1.2}$$

or

$$F(x) = \sum_{j:x_j \le x} p(x_j), \tag{A.1.3}$$

where $\{x_0, x_1, x_2, \ldots\}$ is a finite or countably infinite set. In the case (A.1.2) we shall say that the random variable X is **continuous**. The function f is called the **probability density function** (pdf) of X and can be found from the relation

$$f(x) = F'(x).$$

P.J. Brockwell, R.A. Davis, *Introduction to Time Series and Forecasting*,
Springer Texts in Statistics, DOI 10.1007/978-3-319-29854-2

In case (A.1.3), the possible values of X are restricted to the set $\{x_0, x_1, \ldots\}$, and we shall say that the random variable X is **discrete**. The function p is called the **probability mass function** (pmf) of X, and F is constant except for upward jumps of size $p(x_j)$ at the points x_j. Thus $p(x_j)$ is the size of the jump in F at x_j, i.e.,

$$p(x_j) = F(x_j) - F(x_j^-) = P[X = x_j],$$

where $F(x_j^-) = \lim_{y \uparrow x_j} F(y)$.

A.1.1 Examples of Continuous Distributions

(a) *The normal distribution with mean μ and variance σ^2.* We say that a random variable X has the normal distribution with mean μ and variance σ^2 (written more concisely as $X \sim N(\mu, \sigma^2)$) if X has the pdf given by

$$n\left(x; \mu, \sigma^2\right) = (2\pi)^{-1/2}\sigma^{-1}e^{-(x-\mu)^2/(2\sigma^2)} \qquad -\infty < x < \infty.$$

It follows then that $Z = (X - \mu)/\sigma \sim N(0, 1)$ and that

$$P[X \le x] = P\left[Z \le \frac{x-\mu}{\sigma}\right] = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

where $\Phi(x) = \int_{-\infty}^{x}(2\pi)^{-1/2}e^{-\frac{1}{2}z^2}\,dz$ is known as the **standard normal distribution function**. The significance of the terms *mean* and *variance* for the parameters μ and σ^2 is explained below (see Example A.1.1).

(b) *The uniform distribution on* $[a, b]$. The pdf of a random variable uniformly distributed on the interval $[a, b]$ is given by

$$u(x; a, b) = \begin{cases} \dfrac{1}{b-a}, & \text{if } a \le x \le b, \\ 0, & \text{otherwise.} \end{cases}$$

(c) *The exponential distribution with parameter λ.* The pdf of an exponentially distributed random variable with parameter $\lambda > 0$ is

$$e(x; \lambda) = \begin{cases} 0, & \text{if } x < 0, \\ \lambda e^{-\lambda x}, & \text{if } x \ge 0. \end{cases}$$

The corresponding distribution function is

$$F(x) = \begin{cases} 0, & \text{if } x < 0, \\ 1 - e^{-\lambda x}, & \text{if } x \ge 0. \end{cases}$$

(d) *The gamma distribution with parameters α and λ.* The pdf of a gamma-distributed random variable is

$$g(x; \alpha, \lambda) = \begin{cases} 0, & \text{if } x < 0, \\ x^{\alpha-1}\lambda^{\alpha}e^{-\lambda x}/\,\Gamma(\alpha), & \text{if } x \ge 0, \end{cases}$$

where the parameters α and λ are both positive and Γ is the gamma function defined as

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx.$$

Note that f is the exponential pdf when $\alpha = 1$ and that when α is a positive integer

$$\Gamma(\alpha) = (\alpha - 1)! \text{ with } 0! \text{ defined to be } 1.$$

(e) *The chi-squared distribution with ν degrees of freedom.* For each positive integer ν, the chi-squared distribution with ν degrees of freedom is defined to be the distribution of the sum

$$X = Z_1^2 + \cdots + Z_\nu^2,$$

where $Z_1, \ldots, Z_\nu$ are independent normally distributed random variables with mean 0 and variance 1. This distribution is the same as the gamma distribution with parameters $\alpha = \nu/2$ and $\lambda = \frac{1}{2}$.

A.1.2 Examples of Discrete Distributions

(f) *The binomial distribution with parameters n and p.* The pmf of a binomially distributed random variable X with parameters n and p is

$$b(j; n, p) = P[X = j] = \binom{n}{j} p^j (1-p)^{n-j}, \quad j = 0, 1, \ldots, n,$$

where n is a positive integer and $0 \le p \le 1$.

(g) *The uniform distribution on* $\{1, 2, \ldots, k\}$. The pmf of a random variable X uniformly distributed on $\{1, 2, \ldots, k\}$ is

$$p(j) = P[X = j] = \frac{1}{k}, \quad j = 1, 2 \ldots, k,$$

where k is a positive integer.

(h) *The Poisson distribution with parameter λ.* A random variable X is said to have a Poisson distribution with parameter $\lambda > 0$ if

$$p(j; \lambda) = P[X = j] = \frac{\lambda^j}{j!} e^{-\lambda}, \quad j = 0, 1, \ldots .$$

We shall see in Example A.1.2 below that λ is the mean of X.

(i) *The negative binomial distribution with parameters α and p.* The random variable X is said to have a negative binomial distribution with parameters $\alpha > 0$ and $p \in [0, 1]$ if it has pmf

$$nb(j; \alpha, p) = \left(\prod_{k=1}^{j} \frac{k - 1 + \alpha}{k} \right) (1-p)^j p^\alpha, \quad j = 0, 1, \ldots,$$

where the product is defined to be 1 if $j = 0$.

Not all random variables can be neatly categorized as either continuous or discrete. For example, consider the time you spend waiting to be served at a checkout counter and suppose that the probability of finding no customers ahead of you is $\frac{1}{2}$. Then the time you spend waiting for service can be expressed as

$$W = \begin{cases} 0, & \text{with probability } \dfrac{1}{2}, \\ W_1, & \text{with probability } \dfrac{1}{2}, \end{cases}$$

where W_1 is a continuous random variable. If the distribution of W_1 is exponential with parameter 1, then the distribution function of W is

$$F(x) = \begin{cases} 0, & \text{if } x < 0, \\ \frac{1}{2} + \frac{1}{2}\left(1 - e^{-x}\right) = 1 - \frac{1}{2}e^{-x}, & \text{if } x \geq 0. \end{cases}$$

This distribution function is neither continuous (since it has a discontinuity at $x = 0$) nor discrete (since it increases continuously for $x > 0$). It is expressible as a *mixture*,

$$F = pF_{\rm d} + (1 - p)F_{\rm c},$$

with $p = \frac{1}{2}$, of a discrete distribution function

$$F_{\rm d} = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0, \end{cases}$$

and a continuous distribution function

$$F_{\rm c} = \begin{cases} 0, & x < 0, \\ 1 - e^{-x}, & x \geq 0. \end{cases}$$

Every distribution function can in fact be expressed in the form

$$F = p_1F_{\rm d} + p_2F_{\rm c} + p_3F_{\rm sc},$$

where $0 \leq p_1, p_2, p_3 \leq 1$, $p_1 + p_2 + p_3 = 1$, $F_{\rm d}$ is discrete, $F_{\rm c}$ is continuous, and $F_{\rm sc}$ is *singular continuous* (continuous but not of the form A.1.2). Distribution functions with a singular continuous component are rarely encountered.

A.1.3 Expectation, Mean, and Variance

The **expectation** of a function g of a random variable X is defined by

$$E\left(g(X)\right) = \int g(x)\, dF(x),$$

where

$$\int g(x)\, dF(x) := \begin{cases} \displaystyle\int_{-\infty}^{\infty} g(x)f(x)\, dx & \text{in the continuous case,} \\ \displaystyle\sum_{j=0}^{\infty} g(x_j)p(x_j) & \text{in the discrete case,} \end{cases}$$

and g is any function such that $E|g(x)| < \infty$. (If F is the mixture $F = pF_{\rm c} + (1-p)F_{\rm d}$, then $E(g(X)) = p\int g(x)\, dF_{\rm c}(x) + (1-p)\int g(x)\, dF_{\rm d}(x)$.) The **mean** and **variance** of X are defined as $\mu = EX$ and $\sigma^2 = E(X - \mu)^2$, respectively. They are evaluated by setting $g(x) = x$ and $g(x) = (x - \mu)^2$ in the definition of $E(g(X))$.

It is clear from the definition that expectation has the **linearity property**

$$E(aX + b) = aE(X) + b$$

for any real constants a and b (provided that $E|X| < \infty$).

Example A.1.1 The Normal Distribution

If X has the normal distribution with pdf $n\left(x; \mu, \sigma^2\right)$ as defined in Example (a) above, then

$$E(X-\mu) = \int_{-\infty}^{\infty} (x-\mu) n\big(x; \mu, \sigma^2\big)\, dx = -\sigma^2 \int_{-\infty}^{\infty} n'\big(x : \mu, \sigma^2\big)\, dx = 0.$$

This shows, with the help of the linearity property of E, that

$$E(X) = \mu,$$

i.e., that the parameter μ *is* in fact the mean of the normal distribution defined in Example (a). Similarly,

$$E(X-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 n\big(x; \mu, \sigma^2\big)\, dx = -\sigma^2 \int_{-\infty}^{\infty} (x-\mu) n'\big(x; \mu, \sigma^2\big)\, dx.$$

Integrating by parts and using the fact that f is a pdf, we find that the variance of X is

$$E(X-\mu)^2 = \sigma^2 \int_{-\infty}^{\infty} n\big(x; \mu, \sigma^2\big)\, dx = \sigma^2.$$

□

Example A.1.2 The Poisson Distribution

The mean of the Poisson distribution with parameter λ (see Example (h) above) is given by

$$\mu = \sum_{j=0}^{\infty} \frac{j\lambda^j}{j!} e^{-\lambda} = \sum_{j=1}^{\infty} \frac{\lambda\lambda^{j-1}}{(j-1)!} e^{-\lambda} = \lambda e^{\lambda} e^{-\lambda} = \lambda.$$

A similar calculation shows that the variance is also equal to λ (see Problem A.2).

□

Remark. Functions and parameters associated with a random variable X will be labeled with the subscript X whenever it is necessary to identify the particular random variable to which they refer. For example, the distribution function, pdf, mean, and variance of X will be written as F_X, f_X, μ_X, and σ_X^2, respectively, whenever it is necessary to distinguish them from the corresponding quantities F_Y, f_Y, μ_Y, and σ_Y^2 associated with a different random variable Y.

A.2 Random Vectors

An n-dimensional random vector is a column vector $\mathbf{X} = (X_1, \ldots, X_n)'$ each of whose components is a random variable. The distribution function F of $\mathbf{X}$, also called the **joint distribution** of $X_1, \ldots, X_n$, is defined by

$$F(x_1, \ldots, x_n) = P[X_1, \le x_1, \ldots, X_n \le x_n] \tag{A.2.1}$$

for all real numbers $x_1, \ldots, x_n$. This can be expressed in a more compact form as

$$F(\mathbf{x}) = P[\mathbf{X} \le \mathbf{x}], \quad \mathbf{x} = (x_1, \ldots, x_n)',$$

for all real vectors $\mathbf{x} = (x_1, \ldots, x_n)'$. The joint distribution of any subcollection $X_{i_1}, \ldots, X_{i_k}$ of these random variables can be obtained from F by setting $x_j = \infty$

in (A.2.1) for all $j \notin \{i_1, \ldots, i_k\}$. In particular, the distributions of X_1 and $(X_1, X_n)'$ are given by

$$F_{X_1}(x_1) = P[X_1 \le x_1] = F(x_1, \infty, \ldots, \infty)$$

and

$$F_{X_1,X_n}(x_1, x_n) = P[X_1 \le x_1, X_n \le x_n] = F(x_1, \infty, \ldots, \infty, x_n).$$

As in the univariate case, a random vector with distribution function F is said to be continuous if F has a density function, i.e., if

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f(y_1, \ldots, y_n)\, dy_1\, dy_2 \cdots dy_n.$$

The probability density of $\mathbf{X}$ is then found from

$$f(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n}.$$

The random vector $\mathbf{X}$ is said to be discrete if there exist real-valued vectors $\mathbf{x}_0, \mathbf{x}_1, \ldots$ and a probability mass function $p(\mathbf{x}_j) = P[\mathbf{X} = \mathbf{x}_j]$ such that

$$\sum_{j=0}^{\infty} p(\mathbf{x}_j) = 1.$$

The expectation of a function g of a random vector $\mathbf{X}$ is defined by

$$E\,(g(\mathbf{X})) = \int g(\mathbf{x})\, dF(\mathbf{x}) = \int g(x_1, \ldots, x_n)\, dF(x_1, \ldots, x_n),$$

where

$$\int g(x_1, \ldots, x_n)\, dF(x_1, \ldots, x_n)$$
$$= \begin{cases} \int \cdots \int g(x_1, \ldots, x_n) f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n, & \text{in the continuous case,} \\ \sum_{j_1} \cdots \sum_{j_n} g(x_{j_1}, \ldots, x_{j_n}) p(x_{j_1}, \ldots, x_{j_n}), & \text{in the discrete case,} \end{cases}$$

and g is any function such that $E|g(\mathbf{X})| < \infty$.

The random variables $X_1, \ldots, X_n$ are said to be **independent** if

$$P[X_1 \le x_1, \ldots, X_n \le x_n] = P[X_1 \le x_1] \cdots P[X_n \le x_n],$$

i.e.,

$$F(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$$

for all real numbers $x_1, \ldots, x_n$. In the continuous and discrete cases, independence is equivalent to the factorization of the joint density function or probability mass function into the product of the respective marginal densities or mass functions, i.e.,

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) \tag{A.2.2}$$

or

$$p(x_1, \ldots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n). \tag{A.2.3}$$

For two random vectors $\mathbf{X} = (X_1, \ldots, X_n)'$ and $\mathbf{Y} = (Y_1, \ldots, Y_m)'$ with joint density function $f_{\mathbf{X},\mathbf{Y}}$, the conditional density of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ is

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \begin{cases} \dfrac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})}, & \text{if } f_{\mathbf{X}}(\mathbf{x}) > 0, \\ f_{\mathbf{Y}}(\mathbf{y}), & \text{if } f_{\mathbf{X}}(\mathbf{x}) = 0. \end{cases}$$

The conditional expectation of $g\,(\mathbf{Y})$ given $\mathbf{X} = \mathbf{x}$ is then

$$E(g(\mathbf{Y})|\mathbf{X} = \mathbf{x}) = \int_{-\infty}^{\infty} g(\mathbf{y})\, f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\, d\mathbf{y}.$$

If $\mathbf{X}$ and $\mathbf{Y}$ are independent, then $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = f_{\mathbf{Y}}(\mathbf{y})$ by (A.2.2), and so the conditional expectation of $g(\mathbf{Y})$ given $\mathbf{X} = \mathbf{x}$ is

$$E(g(\mathbf{Y})|\mathbf{X} = \mathbf{x}) = E(\,g(\mathbf{Y})),$$

which, as expected, does not depend on $\mathbf{x}$. The same ideas hold in the discrete case with the probability mass function assuming the role of the density function.

A.2.1 Means and Covariances

If $E|X_i| < \infty$ for each i, then we define the mean or expected value of $\mathbf{X} = (X_1, \ldots, X_n)'$ to be the column vector

$$\boldsymbol{\mu}_X = E\mathbf{X} = (EX_1, \ldots, EX_n)'.$$

In the same way we define the expected value of any array whose elements are random variables (e.g., a matrix of random variables) to be the same array with each random variable replaced by its expected value (if the expectation exists).

If $\mathbf{X} = (X_1, \ldots, X_n)'$ and $\mathbf{Y} = (Y_1, \ldots, Y_m)'$ are random vectors such that each X_i and Y_j has a finite variance, then the **covariance matrix** of $\mathbf{X}$ and $\mathbf{Y}$ is defined to be the matrix

$$\begin{aligned}\Sigma_{\mathbf{XY}} &= \text{Cov}(\mathbf{X}, \mathbf{Y}) = E[(\mathbf{X} - E\mathbf{X})(\mathbf{Y} - E\mathbf{Y})'] \\ &= E(\mathbf{XY}') - (E\mathbf{X})(E\mathbf{Y})'.\end{aligned}$$

The (i, j) element of $\Sigma_{\mathbf{XY}}$ is the covariance $\text{Cov}(X_i, Y_j) = E(X_iY_j) - E(X_i)E(Y_j)$. In the special case where $\mathbf{Y} = \mathbf{X}$, $\text{Cov}(\mathbf{X}, \mathbf{Y})$ reduces to the covariance matrix of the random vector $\mathbf{X}$.

Now suppose that $\mathbf{Y}$ and $\mathbf{X}$ are linearly related through the equation

$$\mathbf{Y} = \mathbf{a} + B\mathbf{X},$$

where $\mathbf{a}$ is an m-dimensional column vector and B is an $m \times n$ matrix. Then $\mathbf{Y}$ has mean

$$E\mathbf{Y} = \mathbf{a} + BE\mathbf{X} \tag{A.2.4}$$

and covariance matrix

$$\Sigma_{\mathbf{YY}} = B\Sigma_{\mathbf{XX}}B' \tag{A.2.5}$$

(see Problem A.3).

Proposition A.2.1 *The covariance matrix $\Sigma_{\mathbf{XX}}$ of a random vector $\mathbf{X}$ is symmetric and nonnegative definite, i.e., $\mathbf{b}'\Sigma_{\mathbf{XX}}\mathbf{b} \ge 0$ for all vectors $\mathbf{b} = (b_1, \ldots, b_n)'$ with real components.*

Proof Since the $(i,\, j)$ element of $\Sigma_{\mathbf{XX}}$ is $\text{Cov}(X_i,\, X_j)$ $=$Cov$(X_j,\, X_i)$, it is clear that $\Sigma_{\mathbf{XX}}$ is symmetric. To prove nonnegative definiteness, let $\mathbf{b} = (b_1, \ldots, b_n)'$ be an arbitrary vector. Then applying (A.2.5) with $\mathbf{a} = \mathbf{0}$ and $B = \mathbf{b}$, we have

$$\mathbf{b}'\Sigma_{\mathbf{XX}}\mathbf{b} = \mathrm{Var}(\mathbf{b}'\mathbf{X}) = \mathrm{Var}(b_1X_1 + \cdots + b_nX_n) \ge 0. \qquad \blacksquare$$

Proposition A.2.2 *Every $n \times n$ covariance matrix Σ can be factorized as*

$$\Sigma = P\Lambda P'$$

where P is an orthogonal matrix (i.e., $P' = P^{-1}$) whose columns are an orthonormal set of right eigenvectors corresponding to the (nonnegative) eigenvalues $\lambda_1, \ldots, \lambda_n$ of Σ, and Λ is the diagonal matrix

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}.$$

In particular, Σ is nonsingular if and only if all the eigenvalues are strictly positive.

Proof Every covariance matrix is symmetric and nonnegative definite by Proposition A.2.1, and for such matrices the specified factorization is a standard result (see Graybill 1983 for a proof). The determinant of an orthogonal matrix is 1 or -1, so that $\det(\Sigma) = \det(P)\det(\Lambda)\det(P) = \lambda_1 \cdots \lambda_n$. It follows that Σ is nonsingular if and only if $\lambda_i > 0$ for all i. ■

Remark 1. Given a covariance matrix Σ, it is sometimes useful to be able to find a square root $A = \Sigma^{1/2}$ with the property that $AA' = \Sigma$. It is clear from Proposition A.2.2 and the orthogonality of P that one such matrix is given by

$$A = \Sigma^{1/2} = P\Lambda^{1/2}P'.$$

If Σ is nonsingular, then we can define

$$\Sigma^s = P\Lambda^s P', \quad -\infty < s < \infty.$$

The matrix $\Sigma^{-1/2}$ defined in this way is then a square root of Σ^{-1} and also the inverse of $\Sigma^{1/2}$. □

A.3 The Multivariate Normal Distribution

The multivariate normal distribution is one of the most commonly encountered and important distributions in statistics. It plays a key role in the modeling of time series data. Let $\mathbf{X} = (X_1, \ldots, X_n)'$ be a random vector.

Definition A.3.1 $\mathbf{X}$ has a **multivariate normal distribution** with mean $\boldsymbol{\mu}$ and nonsingular covariance matrix $\Sigma = \Sigma_{\mathbf{XX}}$, written as $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$, if

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-n/2}(\det\Sigma)^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

If $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$, we can define a *standardized* random vector $\mathbf{Z}$ by applying the linear transformation

$$\mathbf{Z} = \Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu}), \tag{A.3.1}$$

where $\Sigma^{-1/2}$ is defined as in the remark of Section A.2. Then by (A.2.4) and (A.2.5), $\mathbf{Z}$ has mean $\mathbf{0}$ and covariance matrix $\Sigma_{\mathbf{ZZ}} = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = I_n$, where I_n is the $n \times n$ identity matrix. Using the change of variables formula for probability densities (see Mood et al. 1974), we find that the probability density of $\mathbf{Z}$ is

$$\begin{aligned}
f_{\mathbf{Z}}(\mathbf{z}) &= (\det \Sigma)^{1/2} f_{\mathbf{X}}\left(\Sigma^{1/2}\mathbf{z} + \boldsymbol{\mu}\right) \\
&= (\det \Sigma)^{1/2}(2\pi)^{-n/2}(\det \Sigma)^{-1/2} \exp\left\{-\frac{1}{2}(\Sigma^{-1/2}\mathbf{z})'\Sigma^{-1}\Sigma^{-1/2}\mathbf{z}\right\} \\
&= (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\mathbf{z}'\mathbf{z}\right\} \\
&= \left((2\pi)^{-1/2} \exp\left\{-\frac{1}{2}z_1^2\right\}\right) \cdots \left((2\pi)^{-1/2} \exp\left\{-\frac{1}{2}z_n^2\right\}\right),
\end{aligned}$$

showing, by (A.2.2), that $Z_1, \ldots, Z_n$ are independent $\mathrm{N}(0, 1)$ random variables. Thus the standardized random vector $\mathbf{Z}$ defined by (A.3.1) has independent standard normal random components. Conversely, given any $n \times 1$ mean vector $\boldsymbol{\mu}$, a nonsingular $n \times n$ covariance matrix Σ, and an $n \times 1$ vector of standard normal random variables, we can construct a normally distributed random vector with mean $\boldsymbol{\mu}$ and covariance matrix Σ by defining

$$\mathbf{X} = \Sigma^{1/2}\mathbf{Z} + \boldsymbol{\mu}. \tag{A.3.2}$$

(See Problem A.4.)

Remark 1. The multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ can be defined, even when Σ is singular, as the distribution of the vector $\mathbf{X}$ in (A.3.2). The **singular multivariate normal distribution** does not have a joint density, since the possible values of $\mathbf{X} - \boldsymbol{\mu}$ are constrained to lie in a subspace of $\mathbb{R}^n$ with dimension equal to $\mathrm{rank}(\Sigma)$. □

Remark 2. If $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$, B is an $m \times n$ matrix, and $\mathbf{a}$ is a real $m \times 1$ vector, then the random vector

$$\mathbf{Y} = \mathbf{a} + B\mathbf{X}$$

is also multivariate normal (see Problem A.5). Note that from (A.2.4) and (A.2.5), $\mathbf{Y}$ has mean $\mathbf{a} + B\boldsymbol{\mu}$ and covariance matrix $B\Sigma B'$. In particular, by taking B to be the row vector $\mathbf{b}' = (b_1, \ldots, b_n)$, we see that any linear combination of the components of a multivariate normal random vector is normal. Thus $\mathbf{b}'\mathbf{X} = b_1X_1 + \cdots + b_nX_n \sim \mathrm{N}(\mathbf{b}'\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{b}'\Sigma_{\mathbf{XX}}\mathbf{b})$. □

Example A.3.1. The Bivariate Normal Distribution

Suppose that $\mathbf{X} = (X_1, X_2)'$ is a bivariate normal random vector with mean $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \quad \sigma > 0, \sigma_2 > 0, \ -1 < \rho < 1. \tag{A.3.3}$$

The parameters σ_1, σ_2, and ρ are the standard deviations and correlation of the components X_1 and X_2. Every nonsingular 2-dimensional covariance matrix can be expressed in the form (A.3.3). The inverse of Σ is

$$\Sigma^{-1} = \left(1-\rho^2\right)^{-1} \begin{bmatrix} \sigma_1^{-2} & -\rho\sigma_1^{-1}\sigma_2^{-1} \\ -\rho\sigma_1^{-1}\sigma_2^{-1} & \sigma_2^{-2} \end{bmatrix},$$

and so the pdf of $\mathbf{X}$ is given by

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \left(2\pi\sigma_1\sigma_2\left(1-\rho^2\right)^{1/2}\right)^{-1} \\ &\quad \times \exp\left\{\frac{-1}{2\left(1-\rho^2\right)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right.\right. \\ &\qquad \left.\left. -2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}. \end{aligned}$$

□

Multivariate normal random vectors have the important property that the conditional distribution of any set of components, given any other set, is again multivariate normal. In the following proposition we shall suppose that the nonsingular normal random vector $\mathbf{X}$ is partitioned into two subvectors

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix}.$$

Correspondingly, we shall write the mean and covariance matrix of $\mathbf{X}$ as

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

where $\boldsymbol{\mu}^{(i)} = E\mathbf{X}^{(i)}$ and $\Sigma_{ij} = E\left(\mathbf{X}^{(i)}-\boldsymbol{\mu}^{(i)}\right)\left(\mathbf{X}^{(j)}-\boldsymbol{\mu}^{(i)}\right)'$.

Proposition A.3.1.

i. *$\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are independent if and only if $\Sigma_{12} = 0$.*

ii. *The conditional distribution of $\mathbf{X}^{(1)}$ given $\mathbf{X}^{(2)} = \mathbf{x}^{(2)}$ is $\mathrm{N}\big(\boldsymbol{\mu}^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}\big(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}\big), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\big)$. In particular,*

$$E\big(\mathbf{X}^{(1)}|\mathbf{X}^{(2)} = \mathbf{x}^{(2)}\big) = \boldsymbol{\mu}^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}\big(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}\big).$$

The proof of this proposition involves routine algebraic manipulations of the multivariate normal density function and is left as an exercise (see Problem A.6).

Example A.3.2. For the bivariate normal random vector $\mathbf{X}$ in Example A.3.1, we immediately deduce from Proposition A.3.1 that X_1 and X_2 are independent if and only if $\rho\sigma_1\sigma_2 = 0$ (or $\rho = 0$, since σ_1 and σ_2 are both positive). The conditional distribution of X_1 given $X_2 = x_2$ is normal with mean

$$E(X_1|X_2 = x_2) = \mu_1 + \rho\sigma_1\sigma_2^{-1}(x_2 - \mu_2)$$

and variance

$$\mathrm{Var}(X_1|X_2 = x_2) = \sigma_1^2\left(1-\rho^2\right).$$

□

Definition A.3.2. $\{X_t\}$ is a **Gaussian time series** if all of its joint distributions are multivariate normal, i.e., if for any collection of integers $i_1, \ldots, i_n$, the random vector $(X_{i_1}, \ldots, X_{i_n})'$ has a multivariate normal distribution.

Remark 3. If $\{X_t\}$ is a Gaussian time series, then all of its joint distributions are completely determined by the mean function $\mu(t) = EX_t$ and the autocovariance function $\kappa(s,t) = \text{Cov}(X_s, X_t)$. If the process also happens to be stationary, then the mean function is constant ($\mu_t = \mu$ for all t) and $\kappa(t+h, t) = \gamma(h)$ for all t. In this case, the joint distribution of $X_1, \ldots, X_n$ is the same as that of $X_{1+h}, \ldots, X_{n+h}$ for all integers h and $n > 0$. Hence for a Gaussian time series strict stationarity is equivalent to weak stationarity (see Section 2.1). □

Problems

A.1 Let X have a negative binomial distribution with parameters α and p, where $\alpha > 0$ and $0 \le p < 1$.

a. Show that the probability generating function of X $\big($defined as $M(s) = E\big(s^X\big)\big)$ is

$$M(s) = p^\alpha(1 - s + sp)^{-\alpha}, \quad 0 \le s \le 1.$$

b. Using the property that $M'(1) = E(X)$ and $M''(1) = E(X^2) - E(X)$, show that

$$E(X) = \alpha(1-p)/p \quad \text{and} \quad \text{Var}(X) = \alpha(1-p)/p^2.$$

A.2 If X has the Poisson distribution with mean λ, show that the variance of X is also λ.

A.3 Use the linearity of the expectation operator for real-valued random variables to establish (A.2.4) and (A.2.5).

A.4 If Σ is an $n \times n$ covariance matrix, $\Sigma^{1/2}$ is the square root of Σ defined in the remark of Section A.2, and $\mathbf{Z}$ is an n-vector whose components are independent normal random variables with mean 0 and variance 1, show that

$$X = \Sigma^{1/2}\mathbf{Z} + \boldsymbol{\mu}$$

is a normally distributed random vector with mean $\boldsymbol{\mu}$ and covariance matrix Σ.

A.5 Show that if $\mathbf{X}$ is an n-dimensional random vector such that $\mathbf{X} \sim \text{N}(\boldsymbol{\mu}, \Sigma)$, B is a real $m \times n$ matrix, and $\mathbf{a}$ is a real-valued m-vector, then

$$\mathbf{Y} = \mathbf{a} + B\mathbf{X}$$

is a multivariate normal random vector. Specify the mean and covariance matrix of $\mathbf{Y}$.

A.6 Prove Proposition A.3.1.

A.7 Suppose that $\mathbf{X} = (X_1, \ldots, X_n)' \sim \text{N}(\mathbf{0}, \Sigma)$ with Σ nonsingular. Using the fact that $\mathbf{Z}$, as defined in (A.3.1), has independent standard normal components, show that $(\mathbf{X} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{X} - \boldsymbol{\mu})$ has the chi-squared distribution with n degrees of freedom (Section A.1, Example (e)).

A.8 Suppose that $\mathbf{X} = (X_1, \ldots, X_n)' \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$ with Σ nonsingular. If A is a symmetric $n \times n$ matrix, show that $E(\mathbf{X}'A\mathbf{X}) = \operatorname{trace}(A\Sigma) + \boldsymbol{\mu}'\Sigma\boldsymbol{\mu}$.

A.9 Suppose that $\{X_t\}$ is a stationary Gaussian time series with mean 0 and autocovariance function $\gamma(h)$. Find $E(X_t|X_s)$ and $\mathrm{Var}(X_t|X_s)$, $s \neq t$.

B Statistical Complements

B.1 Least Squares Estimation

Consider the problem of finding the "best" straight line

$$y = \theta_0 + \theta_1 x$$

to approximate observations $y_1, \ldots, y_n$ of a dependent variable y taken at fixed values $x_1, \ldots, x_n$ of the independent variable x. The **(ordinary) least squares estimates** $\hat{\theta}_0$, $\hat{\theta}_1$ are defined to be values of θ_0, θ_1 that minimize the sum

$$S(\theta_0, \theta_1) = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

of squared deviations of the observations y_i from the fitted values $\theta_0 + \theta_1 x_i$. (The "sum of squares" $S(\theta_0, \theta_1)$ is identical to the Euclidean squared distance between $\mathbf{y}$ and $\theta_0\mathbf{1} + \theta_1\mathbf{x}$, i.e.,

$$S(\theta_0, \theta_1) = \|\mathbf{y} - \theta_0\mathbf{1} - \theta_1\mathbf{x}\|^2,$$

where $\mathbf{x} = (x_1, \ldots, x_n)'$, $\mathbf{1} = (1, \ldots, 1)'$, and $\mathbf{y} = (y_1, \ldots, y_n)'$.) Setting the partial derivatives of S with respect to θ_0 and θ_1 both equal to zero shows that the vector $\hat{\boldsymbol{\theta}} = (\hat{\theta}_0, \hat{\theta}_1)'$ satisfies the "normal equations"

$$X'X\hat{\boldsymbol{\theta}} = X'\mathbf{y},$$

P.J. Brockwell, R.A. Davis, *Introduction to Time Series and Forecasting*,
Springer Texts in Statistics, DOI 10.1007/978-3-319-29854-2

where X is the $n \times 2$ matrix $X = [\mathbf{1}, \mathbf{x}]$. Since $0 \le S(\boldsymbol{\theta})$ and $S(\boldsymbol{\theta}) \to \infty$ as $\|\boldsymbol{\theta}\| \to \infty$, the normal equations have at least one solution. If $\hat{\boldsymbol{\theta}}^{(1)}$ and $\hat{\boldsymbol{\theta}}^{(2)}$ are two solutions of the normal equations, then a simple calculation shows that

$$\left(\hat{\boldsymbol{\theta}}^{(1)} - \hat{\boldsymbol{\theta}}^{(2)}\right)' X'X \left(\hat{\boldsymbol{\theta}}^{(1)} - \hat{\boldsymbol{\theta}}^{(2)}\right) = 0,$$

i.e., that $X\hat{\boldsymbol{\theta}}^{(1)} = X\hat{\boldsymbol{\theta}}^{(2)}$. The solution of the normal equations is unique if and only if the matrix $X'X$ is nonsingular. But the preceding calculations show that even if $X'X$ is singular, the vector $\hat{\mathbf{y}} = X\hat{\boldsymbol{\theta}}$ of fitted values is the same for *any* solution $\hat{\boldsymbol{\theta}}$ of the normal equations.

The argument just given applies equally well to least squares estimation for the general linear model. Given a set of data points

$$(x_{i1}, x_{i2}, \ldots, x_{im}, y_i), \qquad i = 1, \ldots, n \text{ with } m \le n,$$

the least squares estimate, $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \ldots, \hat{\theta}_m\right)'$ of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ minimizes

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i - \theta_1 x_{i1} - \cdots - \theta_m x_{im})^2 = \left\|\mathbf{y} - \theta_1\mathbf{x}^{(1)} - \cdots - \theta_m\mathbf{x}^{(m)}\right\|^2,$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$ and $\mathbf{x}^{(j)} = (x_{1j}, \ldots, x_{nj})'$, $j = 1, \ldots, m$. As in the previous special case, $\hat{\boldsymbol{\theta}}$ satisfies the equations

$$X'X\hat{\boldsymbol{\theta}} = X'\mathbf{y},$$

where X is the $n \times m$ matrix $X = \left[\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\right]$. The solution of this equation is unique if and only if $X'X$ nonsingular, in which case

$$\hat{\boldsymbol{\theta}} = (X'X)^{-1}X'\mathbf{y}.$$

If $X'X$ is singular, there are infinitely many solutions $\hat{\boldsymbol{\theta}}$, but the vector of fitted values $X\hat{\boldsymbol{\theta}}$ is the same for all of them.

Example B.1.1. To illustrate the general case, let us fit a quadratic function

$$y = \theta_0 + \theta_1 x + \theta_2 x^2$$

to the data

x	0	1	2	3	4
y	1	0	3	5	8

The matrix X for this problem is

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix}, \text{ giving } (X'X)^{-1} = \frac{1}{140}\begin{bmatrix} 124 & -108 & 20 \\ -108 & 174 & -40 \\ 20 & -40 & 10 \end{bmatrix}.$$

The least squares estimate $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2\right)'$ is therefore unique and given by

$$\hat{\boldsymbol{\theta}} = (X'X)^{-1}X'\mathbf{y} = \begin{bmatrix} 0.6 \\ -0.1 \\ 0.5 \end{bmatrix}.$$

The vector of fitted values is given by

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\theta}} = (0.6, 1, 2.4, 4.8, 8.2)'$$

as compared with the observed values

$$\mathbf{y} = (1, 0, 3, 5, 8)'.$$

□

B.1.1 The Gauss–Markov Theorem

Suppose now that the observations $y_1, \ldots, y_n$ are realized values of random variables $Y_1, \ldots, Y_n$ satisfying

$$Y_i = \theta_1 x_{i1} + \cdots + \theta_m x_{im} + Z_i,$$

where $Z_i \sim \mathrm{WN}\big(0, \sigma^2\big)$. Letting $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ and $\mathbf{Z} = (Z_1, \ldots, Z_n)'$, we can write these equations as

$$\mathbf{Y} = X\boldsymbol{\theta} + \mathbf{Z}.$$

Assume for simplicity that the matrix $X'X$ is nonsingular (for the general case see, e.g., Silvey 1975). Then the least squares estimator of $\boldsymbol{\theta}$ is, as above,

$$\hat{\boldsymbol{\theta}} = (X'X)^{-1}X'\mathbf{Y},$$

and the least squares estimator of the parameter σ^2 is the unbiased estimator

$$\hat{\sigma}^2 = \frac{1}{n-m}\big\|\mathbf{Y} - X\hat{\boldsymbol{\theta}}\big\|^2.$$

It is easy to see that $\hat{\boldsymbol{\theta}}$ is also unbiased, i.e., that

$$E\big(\hat{\boldsymbol{\theta}}\big) = \boldsymbol{\theta}.$$

It follows at once that if $\mathbf{c}'\boldsymbol{\theta}$ is any linear combination of the parameters θ_i, $i = 1, \ldots, m$, then $\mathbf{c}'\hat{\boldsymbol{\theta}}$ is an unbiased estimator of $\mathbf{c}'\boldsymbol{\theta}$. The Gauss–Markov theorem says that of all unbiased estimators of $\mathbf{c}'\boldsymbol{\theta}$ of the form $\sum_{i=1}^{n} a_i Y_i$, the estimator $\mathbf{c}'\hat{\boldsymbol{\theta}}$ has the smallest variance.

In the special case where $Z_1, \ldots, Z_n$ are IID $\mathrm{N}\big(0, \sigma^2\big)$, the least squares estimator $\hat{\boldsymbol{\theta}}$ has the distribution $\mathrm{N}\big(\boldsymbol{\theta}, \sigma^2(X'X)^{-1}\big)$, and $(n-m)\hat{\sigma}^2/\sigma^2$ has the χ^2 distribution with $n-m$ degrees of freedom.

B.1.2 Generalized Least Squares

The Gauss–Markov theorem depends on the assumption that the errors $Z_1, \ldots, Z_n$ are uncorrelated with constant variance. If, on the other hand, $\mathbf{Z} = (Z_1, \ldots, Z_n)'$ has mean $\mathbf{0}$ and nonsingular covariance matrix $\sigma^2\Sigma$ where $\Sigma \neq I$, we consider the transformed observation vector $U = R^{-1}\mathbf{Y}$, where R is a nonsingular matrix such that $RR' = \Sigma$. Then

$$\mathbf{U} = R^{-1}X\boldsymbol{\theta} + \mathbf{W} = M\boldsymbol{\theta} + \mathbf{W},$$

where $M = R^{-1}X$ and $\mathbf{W}$ has mean $\mathbf{0}$ and covariance matrix $\sigma^2 I$. The Gauss–Markov theorem now implies that the best linear estimate of any linear combination $\mathbf{c}'\boldsymbol{\theta}$ is $\mathbf{c}'\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}$ is the **generalized least squares estimator**, which minimizes

$$\|\mathbf{U} - M\boldsymbol{\theta}\|^2.$$

In the special case where $Z_1, \ldots, Z_n$ are uncorrelated and Z_i has mean 0 and variance $\sigma^2 r_i^2$, the generalized least squares estimator minimizes the weighted sum of squares

$$\sum_{i=1}^{n} \frac{1}{r_i^2}(Y_i - \theta_1 x_{i1} - \cdots - \theta_m x_{im})^2.$$

In the general case, if $X'X$ and Σ are both nonsingular, the generalized least squares estimator is given by

$$\hat{\boldsymbol{\theta}} = (M'M)^{-1}M'\mathbf{U}.$$

Although the least squares estimator $(X'X)^{-1}X'\mathbf{Y}$ is unbiased if $E(\mathbf{Z}) = \mathbf{0}$, even when the covariance matrix of $\mathbf{Z}$ is not equal to $\sigma^2 I$, the variance of the corresponding estimate of any linear combination of $\theta_1, \ldots, \theta_m$ is greater than or equal to the estimator based on the generalized least squares estimator.

B.2 Maximum Likelihood Estimation

The method of least squares has an appealing intuitive interpretation. Its application depends on knowledge only of the means and covariances of the observations. Maximum likelihood estimation depends on the assumption of a particular distributional form for the observations, known apart from the values of parameters $\theta_1, \ldots, \theta_m$. We can regard the estimation problem as that of selecting the most appropriate value of a parameter vector $\boldsymbol{\theta}$, taking values in a subset Θ of $\mathbb{R}^m$. We suppose that these distributions have probability densities $p(\mathbf{x}; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. For a fixed vector of observations $\mathbf{x}$, the function $L(\boldsymbol{\theta}) = p(\mathbf{x}; \boldsymbol{\theta})$ on Θ is called the **likelihood function**. A maximum likelihood estimate $\hat{\boldsymbol{\theta}}(\mathbf{x})$ of $\boldsymbol{\theta}$ is a value of $\boldsymbol{\theta} \in \Theta$ that maximizes the value of $L(\boldsymbol{\theta})$ for the given observed value $\mathbf{x}$, i.e.,

$$L(\hat{\boldsymbol{\theta}}) = p\big(x; \hat{\boldsymbol{\theta}}(\mathbf{x})\big) = \max_{\boldsymbol{\theta} \in \Theta} p(\mathbf{x}; \boldsymbol{\theta}).$$

Example B.2.1. If $\mathbf{x} = (x_1, \ldots, x_n)'$ is a vector of observations of independent $\mathrm{N}(\mu, \sigma^2)$ random variables, the likelihood function is

$$L\left(\mu, \sigma^2\right) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right], \quad -\infty < \mu < \infty, \quad \sigma > 0.$$

Maximization of L with respect to μ and σ is equivalent to minimization of

$$-2\ln L\left(\mu, \sigma^2\right) = n\ln(2\pi) + 2n\ln(\sigma) + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Setting the partial derivatives of $-2\ln L$ with respect to μ and σ both equal to zero gives the maximum likelihood estimates

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

□

B.2.1 Properties of Maximum Likelihood Estimators

The Gauss–Markov theorem lent support to the use of least squares estimation by showing its property of minimum variance among unbiased linear estimators. Maximum likelihood estimators are not generally unbiased, but in particular cases they can be shown to have small mean squared error relative to other competing estimators. Their main justification, however, lies in their good large-sample behavior.

For independent and identically distributed observations with true probability density $p(\cdot;\boldsymbol{\theta}_0)$ satisfying certain regularity conditions, it can be shown that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ converges in probability to $\boldsymbol{\theta}_0$ and that the distribution of $\sqrt{n}\big(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\big)$ is approximately normal with mean 0 and covariance matrix $I(\boldsymbol{\theta}_0)^{-1}$, where $I(\boldsymbol{\theta})$ is Fisher's information matrix with (i,j) component

$$E_{\boldsymbol{\theta}}\left[\frac{\partial \ln p(X;\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial \ln p(X;\boldsymbol{\theta})}{\partial\theta_j}\right].$$

In time series analysis the situation is rather more complicated than in the case of iid observations. "Likelihood" in the time series context is almost always used in the sense of Gaussian likelihood, i.e., the likelihood computed under the (possibly false) assumption that the series is Gaussian. Nevertheless, estimators of ARMA coefficients computed by maximization of the Gaussian likelihood have good large-sample properties analogous to those described in the preceding paragraph. For details see Brockwell and Davis (1991), Section 10.8.

B.3 Confidence Intervals

Estimation of a parameter or parameter vector by least squares or maximum likelihood leads to a particular value, often referred to as a **point estimate**. It is clear that this will rarely be exactly equal to the true value, and so it is important to convey some idea of the probable accuracy of the estimator. This can be done using the notion of confidence interval, which specifies a random set covering the true parameter value with some specified (high) probability.

Example B.3.1. If $\mathbf{X} = (X_1,\ldots,X_n)'$ is a vector of independent $\mathrm{N}\big(\mu,\sigma^2\big)$ random variables, we saw in Section B.2 that the random variable $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is the maximum likelihood estimator of μ. This is a point estimator of μ. To construct a confidence interval for μ from $\overline{X}_n$, we observe that the random variable

$$\frac{\overline{X}_n-\mu}{S/\sqrt{n}}$$

has Student's t-distribution with $n-1$ degrees of freedom, where S is the sample standard deviation, i.e., $S^2 = \frac{1}{n-1}\sum_{i=1}^n \big(X_i-\overline{X}_n\big)^2$. Hence,

$$P\left[-t_{1-\alpha/2} < \frac{\overline{X}_n-\mu}{S/\sqrt{n}} < t_{1-\alpha/2}\right] = 1-\alpha,$$

where $t_{1-\alpha/2}$ denotes the $(1-\alpha/2)$ quantile of the t-distribution with $n-1$ degrees of freedom. This probability statement can be expressed in the form

$$P\left[\,\overline{X}_n - t_{1-\alpha/2}S/\sqrt{n} < \mu < \overline{X}_n + t_{1-\alpha/2}S/\sqrt{n}\,\right] = 1-\alpha,$$

which shows that the random interval bounded by $\overline{X}_n \pm t_{1-\alpha/2}S/\sqrt{n}$ includes the true value μ with probability $1-\alpha$. This interval is called a $(1-\alpha)$ confidence interval for the mean μ.

□

B.3.1 Large-Sample Confidence Regions

Many estimators of a vector-valued parameter $\boldsymbol{\theta}$ are approximately normally distributed when the sample size n is large. For example, under mild regularity conditions, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}(\mathbf{X})$ of $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_m)'$ is approximately $\mathrm{N}\big(\mathbf{0}, \frac{1}{n}I(\hat{\boldsymbol{\theta}})^{-1}\big)$, where $I(\boldsymbol{\theta})$ is the Fisher information defined in Section B.2. Consequently,

$$n\big(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\big)'I\big(\hat{\boldsymbol{\theta}}\big)\big(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\big)$$

is approximately distributed as χ^2 with m degrees of freedom, and the random set of $\boldsymbol{\theta}$-values defined by

$$n\big(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\big)'I\big(\hat{\boldsymbol{\theta}}\big)\big(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\big) \le \chi^2_{1-\alpha}(m)$$

covers the true value of $\boldsymbol{\theta}$ with probability approximately equal to $1-\alpha$.

Example B.3.2. For iid observations $X_1,\ldots,X_n$ from $\mathrm{N}\big(\mu,\sigma^2\big)$, a straightforward calculation gives, for $\boldsymbol{\theta} = \big(\mu,\sigma^2\big)'$,

$$I(\theta) = \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & \sigma^{-4}/2 \end{bmatrix}.$$

Thus we obtain the large-sample confidence region for $\big(\mu,\sigma^2\big)'$,

$$n\left(\mu-\overline{X}_n\right)^2/\hat{\sigma}^2 + n(\sigma^2-\hat{\sigma}^2)^2/\left(2\hat{\sigma}^4\right) \le \chi^2_{1-\alpha}(2),$$

which covers the true value of $\boldsymbol{\theta}$ with probability approximately equal to $1-\alpha$. This region is an ellipse centered at $\big(\overline{X}_n, \hat{\sigma}^2\big)$.

□

B.4 Hypothesis Testing

Parameter estimation can be regarded as choosing one from infinitely many possible decisions regarding the value of a parameter vector $\boldsymbol{\theta}$. Hypothesis testing, on the other hand, involves a choice between two alternative hypotheses, a “null” hypothesis H_0 and an “alternative” hypothesis H_1, regarding the parameter vector $\boldsymbol{\theta}$. The hypotheses H_0 and H_1 correspond to subsets Θ_0 and Θ_1 of the parameter set Θ. The problem is to decide, on the basis of an observed data vector $\mathbf{X}$, whether or not we should reject the null hypothesis H_0. A statistical test of H_0 can therefore be regarded as a partition of the *sample* space into one set of values of $\mathbf{X}$ for which we reject H_0 and another for which we do not. The problem is to specify a test (i.e., a subset of the sample space called the “rejection region”) for which the corresponding decision rule performs well in practice.

Example B.4.1. If $\mathbf{X} = (X_1, \ldots, X_n)'$ is a vector of independent $\mathrm{N}(\mu, 1)$ random variables, we may wish to test the null hypothesis H_0: $\mu = 0$ against the alternative H_1: $\mu \neq 0$. A plausible choice of rejection region in this case is the set of all samples $\mathbf{X}$ for which $\left|\overline{X}_n\right| > c$ for some suitably chosen constant c. We shall return to this example after considering those factors that should be taken into account in the systematic selection of a "good" rejection region.

□

B.4.1 Error Probabilities

There are two types of error that may be incurred in the application of a statistical test:

- type I error is the rejection of H_0 when it is true.
- type II error is the acceptance of H_0 when it is false.

For a given test (i.e., for a given rejection region R), the probabilities of error can both be found from the **power function** of the test, defined as

$$P_{\boldsymbol{\theta}}(R), \quad \boldsymbol{\theta} \in \Theta,$$

where $P_{\boldsymbol{\theta}}$ is the distribution of $\mathbf{X}$ when the true parameter value is $\boldsymbol{\theta}$. The probabilities of a type I error are

$$\alpha(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(R), \quad \boldsymbol{\theta} \in \Theta_0,$$

and the probabilities of a type II error are

$$\beta(\boldsymbol{\theta}) = 1 - P_{\boldsymbol{\theta}}(R), \quad \boldsymbol{\theta} \in \Theta_1.$$

It is not generally possible to find a test that simultaneously minimizes $\alpha(\boldsymbol{\theta})$ and $\beta(\boldsymbol{\theta})$ for all values of their arguments. Instead, therefore, we seek to limit the probability of type I error and then, subject to this constraint, to minimize the probability of type II error uniformly on Θ_1. Given a **significance level** α, an optimum level-α test is a test satisfying

$$\alpha(\boldsymbol{\theta}) \leq \alpha, \quad \text{for all } \boldsymbol{\theta} \in \Theta_0,$$

that minimizes $\beta(\boldsymbol{\theta})$ for every $\boldsymbol{\theta} \in \Theta_1$. Such a test is called a **uniformly most powerful (U.M.P.) test of level $\boldsymbol{\alpha}$**. The quantity $\sup_{\boldsymbol{\theta} \in \Theta_0} \alpha(\boldsymbol{\theta})$ is called the **size** of the test.

In the special case of a simple hypothesis vs. a simple hypothesis, e.g., H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs. H_1: $\boldsymbol{\theta} = \boldsymbol{\theta}_1$, an optimal test based on the likelihood ratio statistic can be constructed (see Silvey 1975). Unfortunately, it is usually not possible to find a uniformly most powerful test of a simple hypothesis against a composite (more than one value of $\boldsymbol{\theta}$) alternative. This problem can sometimes be solved by searching for uniformly most powerful tests within the smaller classes of unbiased or invariant tests. For further information see Lehmann (1986).

B.4.2 Large-Sample Tests Based on Confidence Regions

There is a natural link between the testing of a simple hypothesis H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs. H_1: $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ and the construction of confidence regions. To illustrate this connection, suppose that $\hat{\boldsymbol{\theta}}$ is an estimator of $\boldsymbol{\theta}$ whose distribution is approximately $\mathrm{N}\big(\boldsymbol{\theta}, n^{-1}I^{-1}(\boldsymbol{\theta})\big)$, where $I(\boldsymbol{\theta})$ is a positive definite matrix. This is usually the case, for example, when $\hat{\boldsymbol{\theta}}$ is a maximum likelihood estimator and $I(\boldsymbol{\theta})$ is the Fisher information.

As in Section B.3.1, we have

$$P_{\boldsymbol{\theta}}\left(n\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)' I\left(\hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right) \le \chi^2_{1-\alpha}(m)\right) \approx 1-\alpha.$$

Consequently, an approximate α-level test is to reject H_0 if

$$n\left(\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}\right)' I\left(\hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}\right) > \chi^2_{1-\alpha}(m),$$

or equivalently, if the confidence region determined by those $\boldsymbol{\theta}$'s satisfying

$$n\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)' I\left(\hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right) \le \chi^2_{1-\alpha}(m)$$

does not include $\boldsymbol{\theta}_0$.

Example B.4.2. Consider again the problem described in Example B.4.1. Since $\overline{X}_n \sim \mathrm{N}\left(\mu, n^{-1}\right)$, the hypothesis H_0: $\mu = 0$ is rejected at level α if

$$n\left(\overline{X}_n\right)^2 > \chi^2_{1-\alpha,1},$$

or equivalently, if

$$\left|\overline{X}_n\right| > \frac{\Phi_{1-\alpha/2}}{n^{1/2}}.$$

□

C Mean Square Convergence

The sequence S_n of random variables is said to converge in mean square to the random variable S if

$$E(S_n - S)^2 \to 0 \text{ as } n \to \infty.$$

In particular, we say that the sum $\sum_{k=1}^{n} X_k$ converges (in mean square) if there exists a random variable S such that $E\left(\sum_{k=1}^{n} X_k - S\right)^2 \to 0$ as $n \to \infty$. If this is the case, then we use the notation $S = \sum_{k=1}^{\infty} X_k$.

C.1 The Cauchy Criterion

For a given sequence S_n of random variables to converge in mean square to *some* random variable, it is necessary and sufficient that

$$E(S_m - S_n)^2 \to 0 \text{ as } m, n \to \infty$$

(for a proof of this see Brockwell and Davis (1991), Chapter 2). The point of the criterion is that it permits checking for mean square convergence without having to identify the limit of the sequence.

Example C.1.1. Consider the sequence of partial sums $S_n = \sum_{t=-n}^{n} a_t Z_t$, $n = 1, 2, \ldots$, where $\{Z_t\} \sim \text{WN}\left(0, \sigma^2\right)$. Under what conditions on the coefficients a_i does this sequence converge in mean square? To answer this question we apply the Cauchy criterion as follows. For $n > m > 0$,

P.J. Brockwell, R.A. Davis, *Introduction to Time Series and Forecasting*,
Springer Texts in Statistics, DOI 10.1007/978-3-319-29854-2

$$E(S_n - S_m)^2 = E\left(\sum_{m<|i|\le n} a_i Z_i\right)^2 = \sigma^2 \sum_{m<|i|\le n} a_i^2 .$$

Consequently, $E(S_n - S_m)^2 \to 0$ if and only if $\sum_{m<|i|\le n} a_i^2 \to 0$. Since the Cauchy criterion applies also to real-valued sequences, this last condition is equivalent to convergence of the sequence $\sum_{i=-n}^{n} a_i^2$, or equivalently to the condition

$$\sum_{i=-\infty}^{\infty} a_i^2 < \infty. \tag{C.1.1}$$

□

Properties of Mean Square Convergence:
If $X_n \to X$ and $Y_n \to Y$, in mean square as $n \to \infty$, then

(a) $E\big(X_n^2\big) \to E\big(X^2\big)$,

(b) $E(X_n) \to E(X)$,

and

(c) $E(X_n Y_n) \to E(XY)$.

Proof. See Brockwell and Davis (1991), Proposition 2.1.2. ■

D Lévy Processes, Brownian Motion and Itô Calculus

D.1 Lévy Processes

Just as ARMA processes were defined as stationary solutions of stochastic difference equations driven by white noise, the so-called CARMA (continuous-time ARMA) models arise as stationary solutions of stochastic differential equations driven by Lévy processes. In order to discuss these equations in more detail we first present a few essential facts concerning Lévy processes. (For detailed accounts see Protter 2010; Applebaum 2004; Bertoin 1996; Sato 1999.) They have already been introduced in Definition 7.5.1, but for ease of reference we repeat the definition here.

Definition D.1.1. A **Lévy process**, $\{L(t), t \in \mathbb{R}\}$ is a process with the following properties:

(i) $L(0) = 0$.

(ii) $L(t) - L(s)$ has the same distribution as $L(t - s)$ for all s and t such that $s \leq t$.

(iii) If (s, t) and (u, v) are disjoint intervals then $L(t) - L(s)$ and $L(v) - L(u)$ are independent.

(iv) $\{L(t)\}$ is continuous in probability, i.e. for all $\epsilon > 0$ and for all $t \in \mathbb{R}$,

$$\lim_{s \to t} P(|L(t) - L(s)| > \epsilon) = 0.$$

P.J. Brockwell, R.A. Davis, *Introduction to Time Series and Forecasting*,
Springer Texts in Statistics, DOI 10.1007/978-3-319-29854-2

It is known that every Lévy process has a version with sample-paths which are right continuous with left limits (càdlàg for short). We shall therefore assume that our Lévy processes have this property.

The characteristic function of $L(t)$, $\phi_t(\theta) := E(\exp(i\theta L(t)))$, has the celebrated Lévy-Khinchin representation, for $t \geq 0$,

$$\phi_t(\theta) = \exp(t\xi(\theta)), \ \theta \in \mathbb{R},$$

where

$$\xi(\theta) = \mathrm{i}\theta\mu - \frac{1}{2}\theta^2\sigma^2 + \int_{\mathbb{R}} (e^{\mathrm{i}\theta x} - 1 - \mathrm{i}\theta x I_{(-1,1)}(x))\nu(dx),$$

for some $\mu \in \mathbb{R}$, $\sigma \geq 0$, and measure ν. $I_{(-1,1)}$ is the indicator function of the set $(-1, 1)$. The measure ν is known as the *Lévy measure* of the process L and satisfies the conditions

$$\nu(\{0\}) = 0$$

and

$$\int_{\mathbb{R}} \min(1, |u|^2)\nu(du) < \infty.$$

The triplet (σ^2, ν, μ) is often referred to as the characteristic triplet of the Lévy process and completely determines all of its finite-dimensional distributions.

The measure ν characterizes the distribution of the jumps of the process. If, in particular, ν is the zero measure then the characteristic function of $L(t)$ for $t \geq 0$, is that of a normal random variable with $E(L(t)) = \mu t$ and $\mathrm{Var}(L(t)) = \sigma^2 t$ and the process $\{L(t), t \in \mathbb{R}\}$ is Brownian motion (Example 7.5.1) with sample-paths which are continuous (but nowhere differentiable).

If $\lambda := \nu(\mathbb{R}) < \infty$ then the expected number of jumps in any time-interval of length t is λt and the expected number of jumps with size in $(-\infty, x]$ in the same time interval is $t\nu((-\infty, x]) = \lambda t F(x)$ where F is a probability distribution function. The distribution function F is known as the jump-size distribution and λ is known as the mean jump-rate. If $\sigma^2 = 0$ and $m = \lambda \int_{(-1,1)} x dF(x)$, then $\{L(t)\}$ is a compound Poisson process with parameters λ and F (Example 7.5.2) and with sample paths which are constant except for jumps.

If $\lambda = \infty$ then the expected number of jumps in every interval of positive length is infinite and the process $\{L(t)\}$ is said to have infinite activity. The gamma process of Example 11.5.1 is such a process with characteristic triplet $(0, \nu, \alpha(1 - e^{-\beta})/\beta)$, where ν is the measure defined on subsets of $(0, \infty)$ by,

$$\nu(dx) = \alpha x^{-1} e^{-\beta x} I_{(0,\infty)}(x) dx.$$

The Lévy-Khinchin representation of the characteristic function of $L(t)$ shows that the distribution of $L(t)$ can, by appropriate choice of the characteristic triplet, be any infinitely divisible distribution. This family includes a vast array of distributions such as the normal distributions, compound Poisson distributions, Student's t-distributions, the stable distributions and many others. In particular it includes distributions which have heavy tails and which are not necessarily symmetric. These features allow for great flexibility when modelling observed phenomena in both financial and physical contexts.

In this appendix we shall restrict attention to Lévy processes for which $EL(1)^2 < \infty$. This constraint is not serious for most applications in finance where it is generally believed that second moments exist while higher moments (those of order four or more) may not. For Lévy processes with $EL(1)^2 < \infty$ it follows from the definition that there are finite constants m and $s \geq 0$ such that

$$EL(t) = mt \text{ and } Var(L(t)) = s^2 t \text{ for all } t \geq 0.$$

In the following sections we shall focus on Brownian motion and stochastic differential equations driven by Brownian motion.

In order to develop the necessary tools we introduce the Itô stochastic integral, Itô processes and Itô's formula. Following this we shall outline some results concerning the solution of stochastic differential equations and use them to expand on the treatment of Gaussian CARMA processes and their Lévy-driven generalizations in Section 11.5.

D.2 Brownian Motion and the Itô Integral

Robert Brown (1828) observed the erratic motion of pollen particles in a liquid which was later explained by the irregular bombardment of the particles by the molecules of the liquid. In order to provide a mathematical model for the one-dimensional version of this process, Einstein (1905) postulated the existence of a process satisfying conditions (i)–(iii) of Definition D.1.1 with $L(t)$ normally distributed for every t. Bachelier (1900) had in fact already proposed such a model for the prices of stocks on the Paris stock exchange. It was later shown by Wiener that there is a process with continuous sample-paths satisfying these conditions, a process which has come to be known as a Brownian motion or Wiener process. It is in fact the only Lévy process with continuous sample-paths, a feature which adds to its plausibility as a model for the physical process originally observed by Brown. Although the sample-paths are continuous they are far from smooth in the sense that they are nowhere differentiable. We shall not attempt to prove these properties here but refer to the books of Mikosch (1998), Klebaner (2005) and Oksendal (2013) for further details. In the following sections we shall give an outline of the essentials of Itô calculus adapted from the more extensive treatment of Øksendal.

For modelling more complex physical phenomena it is often appropriate to suppose that the increment $dX(t)$ of the observed process $\{X(t)\}$ in the infinitesimally small time interval $(t, t + dt)$ satisfies an equation of the form

$$dX(t) = b(t, X(t))dt + \sigma(t, X(t))dB(t), \quad S \leq t \leq T, \tag{D.2.1}$$

where $dB(t)$ denotes the increment of a standard Brownian motion in the same time interval. In order to attach a precise meaning to (D.2.1) we first consider the following discrete approximation. For any fixed positive integer n, consider the grid of time points $\{2^{-n}k, k \in \mathbb{Z}\}$ and define

$$t_k = \begin{cases} 2^{-n}k, & \text{if } S \leq 2^{-n}k \leq T, \\ S, & \text{if } 2^{-n}k < S, \\ T, & \text{if } 2^{-n}k > T. \end{cases} \tag{D.2.2}$$

A discrete approximation to (D.2.1) is then

$$X_{j+1}^n = X_j^n + b(t_j, X_j^n)\Delta t_j + \sigma(t_j, X_j^n)\Delta B_j, \quad [2^n S] \leq j \leq [2^n T], \tag{D.2.3}$$

where $X_j^n := X(t_j)$, $\Delta t_j := t_{j+1} - t_j$, and $\Delta B_j := B(t_{j+1}) - B(t_j)$. For given functions b and σ and for any given initial condition, $X_{[2^n S]}^n = X(S)$, and values of $B(t_j), j \leq k$, equation (D.2.3) can be solved recursively for $X_j^n, j \leq k$. The solution satisfies

$$X_{j+1}^n = X(S) + \sum_{k \leq j} b(t_k, X_k^n)\Delta t_k + \sum_{k \leq j} \sigma(t_k, X_k^n)\Delta B_k, \quad [2^n S] \leq j \leq [2^n T]. \tag{D.2.4}$$

This suggests that, under suitable conditions, as $n \to \infty$, the random variables X_j^n, $[2^n S] \leq j \leq [2^n T] + 1$, approximate (in a sense to be specified) a random process $\{X(t), S \leq t \leq T\}$ satisfying

$$X(t) = X(S) + \int_S^t b(u, X(u))du + \int_S^t \sigma(u, X(u))dB(u), \quad S \leq t \leq T, \tag{D.2.5}$$

In order to make sense of these statements, and to solve equations of the form (D.2.5) we must first define what is meant by the integrals on the right-hand side. We shall do this for **non-anticipating integrands**. The random process $\{X(t)\}$ is said to be a non-anticipating function of $\{B(t)\}$ if, for each t, $X(t)$ is a function of $\{B(s), \ s \leq t\}$. This property is the continuous-time analogue of causality, which we introduced in connection with ARMA processes in Chapter 3. We shall use the notation $\mathscr{F}_t$ to denote the class of random variables on $(\Omega, \mathscr{F}, P)$ (the probability space on which $\{B(t)\}$ is defined) which are functions of $\{B(s), s \leq t\}$. In this terminology $\{X(t)\}$ is a non-anticipating function of $\{B(t)\}$ if $X(t) \in \mathscr{F}_t$ for all t.

To deal with the first integral in (D.2.5) we consider integrals of the form

$$\int_S^T m(u)du, \ S < T, \tag{D.2.6}$$

for functions m on $\mathbb{R} \times \Omega$ belonging to the family $\mathscr{M}(S, T)$ defined by the properties (i)–(iii) below. For clarity we have suppressed the dependence on $\omega \in \Omega$ in (D.2.5) and (D.2.6), but in fact X and m are both functions on $\mathbb{R} \times \Omega$ with values $X(u, \omega)$ and $m(u, \omega)$ respectively.

Defining properties of $m \in \mathscr{M}(S, T)$**:**

(i) $m(\cdot, \cdot)$ is a measurable function on $\mathbb{R} \times \Omega$.

(ii) $m(t, \cdot) \in \mathscr{F}_t$ for each $t \in \mathbb{R}$.

(iii) $P\left[\int_S^T |m(u, \omega)|du < \infty\right] = 1.$

For $m \in \mathscr{M}(S, T)$ the integrals $\int_S^t m(u)du, t \in [S, T]$, can be defined for all ω outside a set of probability zero as straightforward Lebesgue integrals, continuous in t. Specifying them to be zero on the exceptional subset of Ω defines $\int_S^t m(u)du, t \in [S, T]$, as a continuous function of t for each ω.

In order to attach a meaning to the second integral in (D.2.5) we need to define integrals of the form

$$\int_S^T f(u)dB(u), \ S < T, \tag{D.2.7}$$

where the random variables $f(u)$, defined on the same probability space $(\Omega, \mathscr{F}, P)$ as $\{B(t)\}$, satisfy the properties (i)–(iii) specified below. We shall denote the class of such functions as $\mathscr{N}(S, T)$ and an integral of the form (D.2.7) as an **Itô integral**.

Defining properties of $f \in \mathscr{N}(S, T)$:

(i) $f(\cdot, \cdot)$ is a measurable function on $\mathbb{R} \times \Omega$.

(ii) $f(t, \cdot) \in \mathscr{F}_t$ for each $t \in \mathbb{R}$.

(iii) $E\left[\int_S^T f(t,\omega)^2 dt\right] < \infty$.

The construction of the integral (D.2.7) is achieved by defining it for **elementary functions** and then extending the definition to all functions $f \in \mathscr{N}(S, T)$. The function e is an elementary function if for some positive integer n,

$$e(u, \omega) = \sum_{j=-\infty}^{\infty} e_j(\omega) I_{(2^{-n}j, 2^{-n}(j+1)]}(u), \quad u \in \mathbb{R}, \omega \in \Omega, \tag{D.2.8}$$

where the random variables e_j belong to $\mathscr{F}_{t_j}$ for all j and the times t_j are defined as in (D.2.2). Since the function $e(u, \omega)$ is independent of u on the interval $(2^{-n}j, 2^{-n}(j+1)]$, and since B increases on that interval by $\Delta B_j := B(t_{j+1}) - B(t_j)$, it is natural to define (suppressing ω as in (D.2.7)),

$$I_{S,T}(e) = \int_S^T e(u) dB(u) := \sum_{j=-\infty}^{\infty} e_j \Delta B_j, \quad S < T. \tag{D.2.9}$$

Proposition D.2.1. *If e is bounded and elementary then*

$$E\left(\int_S^T e(u) dB(u)\right)^2 = E\left(\int_S^T e(u)^2 du\right), \quad S < T. \tag{D.2.10}$$

Proof. Observing that $E(e_i e_j \Delta B_i \Delta B_j) = \delta_{ij} E(e_j^2) \Delta t_j$, where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise, we can rewrite the left-hand side of (D.2.10) as

$$E \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} (e_i e_j \Delta B_i \Delta B_j) = \sum_{j=-\infty}^{\infty} E(e_j^2) \Delta t_j = E \sum_{j=-\infty}^{\infty} e_j^2 \Delta t_j = E \int_S^T e(t)^2 \, dt.$$

■

Remark 1. The left-hand side of (D.2.10) is the squared norm of the random variable $I_{S,T}(e)$ defined on $(\Omega, \mathscr{F}, P)$. The right-hand side is the squared norm of the function

$$e^*(u, \omega) := \begin{cases} e(u, \omega), & \text{if } (u, \omega) \in [S, T] \times \Omega, \\ 0, & \text{otherwise}, \end{cases}$$

a square integrable function on the product space $[S, T] \times \Omega$ with respect to the product measure $\ell \times P$, where ℓ denotes Lebesgue measure. The mapping $e \mapsto I_{S,T}(e)$ thus determines an **isometry** from the restrictions e^* of the bounded elementary functions e to $[S, T] \times \Omega$ into the space of square integrable random variables on $(\Omega, \mathscr{F}, P)$.

It can be shown (see e.g., Oksendal 2013) that for every function $f \in \mathscr{N}(S, T)$ there is a sequence of bounded elementary functions $\{e_n\}$ such that

$$E \int_S^T (e_n(u) - f(u))^2 du \to 0 \text{ as } n \to \infty. \tag{D.2.11}$$

This implies that $E\int_S^T(e_n(u)-e_m(u))^2du \to 0$ as m and n both go to ∞ and, by the isometry, that

$$E(I_{S,T}(e_n-e_m))^2 = E\left(I_{S,T}(e_n)-I_{S,T}(e_m)\right)^2 \to 0.$$

By the Cauchy property of mean square convergence (Appendix C.1) it follows that $\{I_{S,T}(e_n)\}$ has a mean square limit.

If $\{g_n\}$ is another sequence of bounded elementary functions with the property (D.2.11) then $E\int_S^T(e_n(u)-g_n(u))^2du \to 0$ as $n\to\infty$ so that

$$E(I_{S,T}((e_n-g_n))^2 = E(I_{S,T}(e_n)-I_{S,T}(g_n))^2 \to 0.$$

Hence the mean square limit of $I_{S,T}(e_n)$ is the same for all sequences of bounded elementary functions satisfying (D.2.11) and the common limit is defined to be $I_{S,T}(f)$. Thus $I_{S,T}(f)$ can be defined unambiguously as

$$I_{S,T}(f) := \lim_{m.s.} I_{S,T}(e_n), \tag{D.2.12}$$

where $\{e_n\}$ is any sequence of bounded elementary functions satisfying (D.2.11).

Moreover if $f\in\mathscr{N}(S,T)$ and $\{e_n\}$ satisfies (D.2.11), then

$$E\left(I_{S,T}(f)^2\right) = \lim_{n\to\infty} E\left(I_{S,T}(e_n)^2\right) = \lim_{n\to\infty} E\int_S^T e_n^2(u)du = E\int_S^T f^2(u)du,$$

showing that the isometry of the restrictions e^* of bounded elementary functions extends to the corresponding restrictions f^* of *all* functions in $\mathscr{N}(S,T)$.

This means that, in principle, $I_{S,T}(f)$ can be evaluated as the mean-square limit of $\int_S^T x_n(u)dB(u)$ where $\{x_n\}$ is *any* (not necessarily bounded) sequence of elementary functions such that $E\int_S^T(x_n(u)-f(u))^2du \to 0$ as $n\to\infty$. In particular it can be shown in this way that

$$\int_S^T B(u)dB(u) = \frac{1}{2}(B^2(T)-B^2(S)) - \frac{1}{2}(T-S).$$

We shall not go into the details as we shall derive this result in a much simpler way using the tools of Itô calculus to be discussed in the following section. □

Remark 2. If $f\in\mathscr{N}(S,T)$ then for each $t\in[S,T]$ so also is the function, $\{f(\omega,u)\mathbf{1}_{[S,t]}(u), \omega\in\Omega, u\in\mathbb{R}\}$, where $\mathbf{1}_{[S,t]}$ is the indicator function of the set $[S,t]$. This enables us to define

$$\int_S^t f(u)dB(u) := \int_S^T f(u)\mathbf{1}_{[S,\,t]}(u)du$$

for each $t\in[S,T]$ and each $f\in\mathscr{N}(S,T)$. □

Remark 3. If $f\in\mathscr{N} := \cap\mathscr{N}(S,T)$, where $\cap$ denotes the intersection over all $S\in\mathbb{R}$ and $T\in\mathbb{R}$ such that $S\le T$, then $I_{s,t}(f)$ is defined for all real-valued s and t such that $s\le t$ and the integral has the properties,

(i) $EI_{s,t}(f)=0$.
(ii) $I_{s,u}(f) = I_{s,t}(f)+I_{t,u}(f),\ s\le t\le u.$
(iii) $I_{s,t}(af+bg) = aI_{s,t}(f)+bI_{s,t}(g)$ for all $a,b\in\mathbb{R}$ and $g\in\mathscr{N}$.
(iv) $E\left[I_{s,t}(f)I_{s,t}(g)\right] = E\int_s^t f(u)g(u)du$ for all $g\in\mathscr{N}$.

(v) For each fixed $s \in \mathbb{R}$, $\{I_{s,t}(f), t \geq s\}$ is an $\mathscr{F}_t$-martingale, i.e. $E|I_{s,t}(f)| < \infty$ and

$$E(I_{s,u}(f)|B(y), y \leq t) = I_{s,t}(f), \;\; u \geq t \geq s.$$

(vi) For each fixed $s \in \mathbb{R}$ and for each fixed $T \geq s$ there is a version of $\{I_{s,t}(f), s \leq t \leq T\}$ which is continuous in t. In other words there is a process $\{X_t, s \leq t \leq T\}$ with continuous sample-paths such that

$$P(X_t = \int_s^t f(u)dB(u)) = 1 \text{ for all } t \in [s, T].$$

Properties (i)–(iv) are clearly true for bounded elementary functions f and g. Their validity for functions in $\mathscr{N}$ can be established by taking limits. Property (v) follows from (ii) and the independence of the increments of $\{B(t)\}$. The proof of property (vi) is beyond the scope of this book [see, e.g., Oksendal (2013) for details]. □

D.3 Itô Processes and Itô's Formula

Direct evaluation of Itô stochastic integrals from the definition (D.2.12) is very messy. For example, it can be shown by a lengthy calculation from the definition that

$$\int_0^t B(u)dB(u) = \frac{1}{2}B(t)^2 - \frac{1}{2}t.$$

Itô's formula provides a chain rule for evaluating such integrals. It is clear from this example that the classic rule for Riemann integration does not apply. If, for example, we apply it in this particular case we find, from the rule $d(x^2) = 2xdx$, that the integral is $\frac{1}{2}B(t)^2$ instead of the correct expression above. Before we can derive the appropriate rule however we first need to define what is meant by an Itô process.

Itô Process

This is a process which satisfies (suppressing the argument ω as before)

$$X(t) = X(s) + \int_s^t m(u)du + \int_s^t f(u)dB(u), \;\; s \leq t \in \mathbb{R}, \tag{D.3.1}$$

where

$$X(t) \in \mathscr{F}_t \text{ for all } t \in \mathbb{R}, \tag{D.3.2}$$

$$m \in \mathscr{M}(S, T) \text{ for all } S \leq T \in \mathbb{R} \tag{D.3.3}$$

and

$$f \in \mathscr{N}^*(S, T) \text{ for all } S \leq T \in \mathbb{R}, \tag{D.3.4}$$

with $\mathscr{M}(S, T)$ defined as in Section E.2 and $\mathscr{N}^*(S, T)$ defined like $\mathscr{N}(S, T)$ in Section E.2 except for the replacement of property (iii), $E\left[\int_S^T f(u)^2du\right] < \infty$, by the weaker condition,

(iii)* $P\left[\int_S^T f(u)^2du < \infty\right] = 1$.

It can be shown that, under this weaker condition, the integrals $I_{s,t}(f)$, $s \leq t \in \mathbb{R}$, can still be defined, retaining all of the properties in Remark 3 of Section E.2 with the exception of the martingale property (v).

Definition (D.3.1) is often written in the shorthand notation,

$$dX(t) = m(t)\,dt + f(t)\,dB(t). \tag{D.3.5}$$

Both of the integrals in (D.3.1) are assumed to be continuous versions so that the Itô process $\{X(t)\}$ is also continuous. The first integral is usually referred to as the **drift** component of $\{X(t)\}$ and the second as the **Brownian** component.

Itô's Formula

Itô's formula is concerned with smooth functions of Itô processes. Specifically it states that if $\{X(t)\}$ is an Itô process satisfying (D.3.5) and $\{g(t,x)\}$ is a function on $\mathbb{R}\times\mathbb{R}$ with continuous partial derivatives $\partial g/\partial t$ and $\partial^2 g/\partial x^2$ then

(i) $Y(t) := g(t, X(t))$ is an Itô process and

(ii)

$$dY(t) = \frac{\partial g}{\partial t}(t,X(t))\,dt + \frac{\partial g}{\partial x}(t,X(t))\,dX(t) + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}(t,X(t))\,(dX(t))^2, \tag{D.3.6}$$

where $dX(t) = m\,dt + f\,dB(t)$ and $(dX(t))^2 = f^2\,dt$.

Writing g_t, g_x and g_{xx} for the corresponding partial derivatives of g evaluated at $(t, X(t))$, and substituting for $dX(t)$ and $dX(t)^2$ as indicated in (ii), we can write the increment of $Y(t)$ explicitly in the form (D.3.5) as

$$dY(t) = (g_t + mg_x + \frac{1}{2}v^2 g_{xx})\,dt + fg_x\,dB(t). \tag{D.3.7}$$

Example D.3.1. $\int_0^t B(u)dB(u)$

Inspection of (D.3.7) suggests that in order to find a process with increments $B(u)dB(u)$ we should start with the Itô process $X(t) = B(t)$, for which $m = 0$ and $f = 1$, and define $Y(t) = g(t, X(t))$ where $g_x(t,x) = x$. Taking $g(t,x) = x^2/2$ we obtain, from (D.3.7),

$$dY(t) = \frac{1}{2}\,dt + B(t)dB(t),$$

which gives

$$\int_0^t B(u)dB(u) = Y(t) - Y(0) - \frac{1}{2}t = \frac{1}{2}B(t)^2 - \frac{1}{2}t.$$

□

Multivariate Itô Processes

An n-dimensional Itô process $\{\mathbf{X}(t)\}$ is defined to be an n-dimensional vector-valued process satisfying an equation (cf. (D.3.5)),

$$d\mathbf{X}(t) = \mathbf{m}(t)\,dt + F(t)\,d\mathbf{B}(t), \tag{D.3.8}$$

where $\{\mathbf{B}(t)\}$ is m-dimensional standard Brownian motion, i.e. an m-dimensional random process with components which are independent one-dimensional standard Brownian motions, the components of the n-vectors $\mathbf{X}(t)$ and $\mathbf{m}(t)$ satisfy (D.3.2) and (D.3.3) respectively, and each component f_{ij} of the $n\times m$ matrix $F(t)$ satisfies (D.3.4). The more explicit form of (D.3.8), corresponding to (D.3.1), is

$$\mathbf{X}(t) = \mathbf{X}(s) + \int_s^t \mathbf{m}(u)\,du + \int_s^t F(u)\,d\mathbf{B}(u),\ \ s \le t \in \mathbb{R}. \tag{D.3.9}$$

The Multidimensional Itô Formula

The multidimensional version of Itô's formula states that if $\{\mathbf{X}(t)\}$ is an n-dimensional Itô process satisfying (D.3.9) and $\{g(t, \mathbf{x})\}$ is a function on $\mathbb{R} \times \mathbb{R}^n$ with values in $\mathbb{R}^p$ and with continuous second partial derivatives, then

(i) $\{\mathbf{Y}(t) := g(t, \mathbf{X}(t))$ is a p-dimensional Itô process and

(ii)

$$dY_i(t) = \frac{\partial g_i}{\partial t}\, dt + \sum_{j=1}^{n} \frac{\partial g_i}{\partial x_j}\, dX_j(t) + \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{\partial^2 g_i}{\partial x_j \partial x_k}\, dX_j(t) dX_k(t), \tag{D.3.10}$$

where X_i, Y_i and g_i are the components of $\mathbf{X}$, $\mathbf{Y}$ and g respectively, and the partial derivatives of g are all evaluated at $(t, \mathbf{X}(t))$. The increments dX_j satisfy the relations $dX_j(t) = m_j\, dt + \sum_{r=1}^{n} f_{jr}\, dB_r(t)$ and $dX_j(t)dX_k(t) = \sum_{r=1}^{n} f_{jr} f_{kr}\, dt$, where m_j and f_{ij} are the components of $\mathbf{m}(t)$ and $F(t)$ respectively.

In the following section we shall consider solutions of stochastic differential equations of the form

$$d\mathbf{X}(t) = b(t, \mathbf{X}(t))\, dt + \sigma(t, \mathbf{X}(t))\, d\mathbf{B}(t), \quad S < t < T, \quad \mathbf{X}_S = Z, \tag{D.3.11}$$

where $\{\mathbf{B}(t)\}$ is m-dimensional standard Brownian motion. Conditions on the functions b and σ and the initial random variable Z which guarantee existence and uniqueness of solutions will be specified in Theorem D.4.1, a proof of which can be found in Oksendal (2013).

D.4 Itô Stochastic Differential Equations

The equation (D.3.11) is known as an *Itô stochastic differential equation* for the $\mathbb{R}^n$-valued random process $\{\mathbf{X}(t)\}$. Equations (7.5.6), for geometric Brownian motion, (11.5.2), for the CAR(1) process, and (11.5.9), for the state vector of a CARMA process, are special cases. It is trivial to check, in each of these cases, that the conditions on b and σ given in the following theorem are satisfied for all S and $T \in \mathbb{R}$ with $T > S$. Provided the conditions on the initial random vector $\mathbf{Z}$ are satisfied, these guarantee the existence and uniqueness of a continuous solution of (D.3.11). After stating the theorem we shall use Itô's formula to derive solutions of the particular Itô equations (7.5.6) and (11.5.9). The solution of (11.5.2) was discussed in Section 11.5.1.

Theorem D.4.1. *Suppose that $S < T \in \mathbb{R}$ and that the measurable functions $b : [S, T] \times \mathbb{R}^n \mapsto \mathbb{R}^n$ and $\sigma : [S, T] \times \mathbb{R}^n \mapsto \mathbb{R}^n \times \mathbb{R}^m$ in (D.3.11) have the properties*

$$|b(t, \mathbf{x})| + |\sigma(t, \mathbf{x})| < C(1 + |\mathbf{x}|), \ \mathbf{x} \in \mathbb{R}^n, \ t \in [S, T]$$

and

$$|b(t, \mathbf{x} - b(t, \mathbf{y})| + |\sigma(t, \mathbf{x} - \sigma(t, \mathbf{y})| < \ D|\mathbf{x} - \mathbf{y}|,$$

where C and D are finite positive constants and $|M|$ denotes the (positive) square root of the sum of squares of the components of the matrix or vector M. If Z is a random variable independent of $\{B(t) - B(s), S \leq s < t \leq T\}$ such that $E|Z|^2 < \infty$, then the *stochastic differential equation (D.3.11) has a unique continuous (in t) solution, each component of which belongs to $\mathcal{N}^*[S, T]$ as defined in (D.3.4).*

Geometric Brownian Motion

Geometric Brownian motion was introduced in Section 7.5.2 as a continuous-time model for asset prices and was the basis for the derivations by Black and Scholes (1973) and Merton (1973) of the option-pricing formula discussed in Section 7.6. Here we shall use Itô's formula to find the solution $\{P(t),\ t \geq 0\}$ of the defining differential equation,

$$dP(t) = P(t)[\mu dt + \sigma dB(t)],\ \ t \geq 0, \tag{D.4.1}$$

where $P(0)$ is a strictly positive random variable, independent of $\{B(t) - B(s), 0 \leq s \leq t < \infty\}$. The standard calculus identity, $d(\log(y)) = dy/y$, suggests that we try applying Itô's formula with $X(t) = P(t)$ and $g(x, t) = \log(x)$. The function g has continuous partial derivatives, $\partial g/\partial t = 0$, $\partial g/\partial x = 1/x$ and $\partial^2 g/\partial x^2 = -1/x^2$ on the set where $x > 0$. Substituting in (D.3.6) and using (D.4.1) we obtain □

$$d(\log P(t)) = \frac{1}{P(t)}\, dP(t) - \frac{1}{2P(t)^2}(dP(t))^2 = \mu dt + \sigma dB(t) - \frac{\sigma^2}{2}\, dt, \tag{D.4.2}$$

whence

$$\log(P(t)) - \log(P(0)) = (\mu - \frac{\sigma^2}{2})t + \sigma B(t).$$

This is equivalent to the solution (7.5.7) given earlier.

Gaussian CARMA Processes

The state equation (11.5.9) for the Gaussian CARMA(p, q) process, i.e.

$$d\mathbf{X}(t) = A\mathbf{X}(t)dt + \mathbf{e}dB(t), \tag{D.4.3}$$

where $\mathbf{X}(0)$ is independent of $\{B(t)-B(s), 0 \leq s \leq t \leq T\}$ and $E|\mathbf{X}(\mathbf{0})|^2 < \infty$, clearly satisfies the conditions of Theorem D.4.1 and therefore has a unique solution which is continuous in t. In order to find the solution we multiply both sides by the integrating factor e^{-At}, as we would if $\{B(t)\}$ were deterministic. Since e^{-At} is non-singular the state equation is equivalent to the equation

$$e^{-At}d\mathbf{X}(t) - e^{-At}A\mathbf{X}(t)dt = e^{-At}\mathbf{e}dB(t). \tag{D.4.4}$$

This form of the equation suggests applying the multivariate Itô formula with $\mathbf{g}(t, \mathbf{x}) = e^{-At}\mathbf{x}$. The second derivatives of g are all continuous and satisfy

$$\frac{\partial^2 g_i}{\partial x_j \partial x_k} = 0 \text{ for all } i, j \text{ and } k,$$

$$\frac{\partial g_i}{\partial x_j} = \mathbf{e}_i' e^{-At}\mathbf{e}_j,$$

and

$$\frac{\partial \mathbf{g}}{\partial t} = -\mathbf{e}_i' A e^{-At}\mathbf{x}.$$

where $\mathbf{e}_r$, $r \in \{1, \ldots, p\}$, denotes a p-component column vector, all of whose components are zero except for the rth, which is one. Substituting these derivatives into (D.3.10) and writing the resulting equations in vector form we obtain

$$d(e^{-At}\mathbf{X}(t)) = -Ae^{-At}\mathbf{X}(t)dt + e^{-At}d\mathbf{X}(t).$$

Substituting this expression in (D.4.4) gives

$$d(e^{-At}\mathbf{X}(t)) = e^{-At}dB(t),$$

which implies that

$$e^{-At}\mathbf{X}(t) - \mathbf{X}(0) = \int_0^t e^{-Au}\mathbf{e}dB(u),$$

or equivalently

$$\mathbf{X}(t) = e^{At}\mathbf{X}(0) + \int_0^t e^{A(t-u)}\mathbf{e}dB(u), \ 0 \le t \le T. \tag{D.4.5}$$

Since equation (D.4.1), with $\mathbf{X}(S)$ independent of $\{B(t) - B(s), S \le s \le t \le T\}$ and $E|\mathbf{X}(\mathbf{S})|^2 < \infty$, satisfies the conditions of Theorem D.4.1 for *all* $S \in \mathbb{R}$ and $T \in \mathbb{R}$ such that $S < T$, exactly the same arguments give the more general relation,

$$\mathbf{X}(t) = e^{A(t-S)}\mathbf{X}(S) + \int_S^t e^{A(t-u)}\mathbf{e}\, dB(u), \ t \ge S, \ \text{for all } S \in \mathbb{R}. \tag{D.4.6}$$

This is equation (11.5.11) for which we showed (in Section 11.5.2) that the unique causal stationary solution is

$$\mathbf{X}(t) = \int_{-\infty}^t e^{A(t-u)}\mathbf{e}\, dB(u), \ t \in \mathbb{R}.$$

This led, with (11.5.8), to the definition of the zero-mean causal CARMA(p, q) process $\{Y(t), t \in \mathbb{R}\}$ as

$$Y(t) = \int_{-\infty}^t \mathbf{b}'e^{A(t-u)}\mathbf{e}\, dB(u)$$

and, more generally in Section 11.5.3, to the second-order Lévy-driven CARMA(p, q) process,

$$Y(t) = \int_{(-\infty,t]} \mathbf{b}'e^{A(t-u)}\mathbf{e}\, dL(u).$$

E An ITSM Tutorial

The package ITSM2000 requires an IBM-compatible PC operating under Windows XP or any subsequent Windows operating system. To install the package, go to http://extras.springer.com and locate the extras for this book either by entering the ISBN number or by choosing the year 2016. Choose the option *Download Entire Contents* and you will receive a zip file containing ITSM.EXE, the data files, an introduction called README.PDF and a searchable document of Help files, ITSM_HELP.PDF. For a quick and easy introduction to the use of the package we recommend following the instructions in README.PDF. For detailed help on each of the functions of the program refer to ITSM_HELP.PDF. Under older Windows operating systems the Help files can be accessed from the Help menu within ITSM itself, but this feature is not yet supported by all versions of Windows so you may need to open ITSM_HELP.PDF in a separate window while running the program.

When you unzip the downloaded zip file it will create a folder called ITSM2000 which contains all the necessary files for running the program. Double-click on the ITSM icon or the ITSM-Shortcut icon to open the ITSM window. (You may wish to copy and paste the ITSM-Shortcut icon to the desktop or some other convenient location from which it can also be accessed.). The package ITSM2000 supersedes the versions of the package ITSM2000 distributed with earlier editions of this book.

P.J. Brockwell, R.A. Davis, *Introduction to Time Series and Forecasting*,
Springer Texts in Statistics, DOI 10.1007/978-3-319-29854-2

E.1 Getting Started

E.1.1 Running ITSM

Double-clicking on the ITSM or the ITSM-Shortcut icon will open the ITSM window. To analyze one of the data sets provided, select `File>Project>Open` at the top left corner of the ITSM window.

There are several distinct functions of the program ITSM. The first is to analyze and display the properties of time series data, the second is to compute and display the properties of time series models, and the third is to combine these functions in order to fit models to data. The last of these includes checking that the properties of the fitted model match those of the data in a suitable sense. Having found an appropriate model, we can (for example) then use it in conjunction with the data to forecast future values of the series. Sections E.2–E.5 of this appendix deal with the modeling and analysis of data, while Section E.6 is concerned with model properties. Section E.7 explains how to open multivariate projects in ITSM. Examples of the analysis of multivariate time series are given in Chapter 8.

It is important to keep in mind the distinction between data and model properties and not to confuse the data with the model. In any one project ITSM stores one data set and one model (which can be identified by highlighting the project window and pressing the red INFO button at the top of the ITSM window). Until a model is entered by the user, ITSM stores the default model of white noise with variance 1. If the data are transformed (e.g., differenced and mean-corrected), then the data are replaced in ITSM by the transformed data. (The original data can, however, be restored by inverting the transformations.) Rarely (if ever) is a real time series generated by a model as simple as those used for fitting purposes. In model fitting the objective is to develop a model that mimics important features of the data, but is still simple enough to be used with relative ease.

The following sections constitute a tutorial that illustrates the use of some of the features of ITSM by leading you through a complete analysis of the well-known airline passenger series of Box and Jenkins (1976) filed as AIRPASS.TSM in the ITSM2000 folder.

E.2 Preparing Your Data for Modeling

The observed values of your time series should be available in a single-column ASCII file (or two columns for a bivariate series). The file, like those provided with the package, should be given a name with suffix .TSM. You can then begin model fitting with ITSM. The program will read your data from the file, plot it on the screen, compute sample statistics, and allow you to make a number of transformations designed to make your transformed data representable as a realization of a zero-mean stationary process.

Example E.2.1. To illustrate the analysis we shall use the file AIRPASS.TSM, which contains the number of international airline passengers (in thousands) for each month from January, 1949, through December, 1960.

□

E.2.1 Entering Data

Once you have opened the ITSM window as described above under Getting Started, select the options `File>Project>Open`, and you will see a dialog box in which you can check either `Univariate` or `Multivariate`. Since the data set for this example is univariate, make sure that the univariate option is checked and then click `OK`. A window labeled `Open File` will then appear, in which you can either type the name AIRPASS.TSM and click `Open`, or else locate the icon for AIRPASS.TSM in the Open File window and double-click on it. You will then see a graph of the monthly international airline passenger totals (measured in thousands) $X_1, \ldots, X_n$, with $n = 144$. Directly behind the graph is a window containing data summary statistics.

An additional, second, project can be opened by repeating the procedure described in the preceding paragraph. Alternatively, the data can be *replaced* in the current project using the option `File>Import File`. This option is useful if you wish to examine how well a fitted model represents a different data set. (See the entry `ProjectEditor` in the ITSM_HELP Files for information on multiple project management. Each ITSM project has its own data set and model.) For the purpose of this introduction we shall open only one project.

E.2.2 Information

If, with the window labeled AIRPASS.TSM highlighted, you press the red INFO button at the top of the ITSM window, you will see the sample mean, sample variance, estimated standard deviation of the sample mean, and the current model (white noise with variance 1).

Example E.2.2. Go through the steps in Entering Data to open the project AIRPASS.TSM and use the INFO button to determine the sample mean and variance of the series.

□

E.2.3 Filing Data

You may wish to transform your data using ITSM and then store it in another file. At any time before or after transforming the data in ITSM, the data can be exported to a file by clicking on the red `Export` button, selecting `Time Series` and `File`, clicking `OK`, and specifying a new file name. The numerical values of the series can also be pasted to the clipboard (and from there into another document) in the same way by choosing `Clipboard` instead of `File`. Other quantities computed by the program (e.g., the residuals from the current model) can be filed or pasted to the clipboard in the same way by making the appropriate selection in the Export dialog box. Graphs can also be pasted to the clipboard by right-clicking on them and selecting `Copy to Clipboard`.

Example E.2.3. Copy the series AIRPASS.TSM to the clipboard, open Wordpad or some convenient screen editor, and choose `Edit>Paste` to insert the series into your new document. Then copy the graph of the series to the clipboard and insert it into your document in the same way.

□

E.2.4 Plotting Data

A time series graph is automatically plotted when you open a data file (with time measured in units of the interval between observations, i.e., $t = 1, 2, 3, \ldots$). To see a histogram of the data press the rightmost yellow button at the top of the ITSM screen. If you wish to adjust the number of bins in the histogram, select `Statistics>Histogram>Set Bin Count` and specify the number of bins required. The histogram will then be replotted accordingly.

To insert any of the ITSM graphs into a text document, right-click on the graph concerned, select `Copy to Clipboard`, and the graph will be copied to the clipboard. It can then be pasted into a document opened by any standard text editor such as MS-Word or Wordpad using the `Edit>Paste` option in the screen editor. The graph can also be sent directly to a printer by right-clicking on the graph and selecting `Print`. Another useful graphics feature is provided by the white Zoom buttons at the top of the ITSM screen. The first and second of these enable you to enlarge a designated segment or box, respectively, of any of the graphs. The third button restores the original graph.

Example E.2.4. Continuing with our analysis of AIRPASS.TSM, press the yellow histogram button to see a histogram of the data. Replot the histogram with 20 bins by selecting `Statistics>Histogram>Set Bin Count`.

□

E.2.5 Transforming Data

Transformations are applied in order to produce data that can be successfully modeled as "stationary time series." In particular, they are used to eliminate trend and cyclic components and to achieve approximate constancy of level and variability with time.

Example E.2.5. The airline passenger data (see Figure 10-4) are clearly not stationary. The level and variability both increase with time, and there appears to be a large seasonal component (with period 12). They must therefore be transformed in order to be represented as a realization of a stationary time series using one or more of the transformations available for this purpose in ITSM.

□

Box–Cox Transformations

Box–Cox transformations are performed by selecting `Transform>Box-Cox` and specifying the value of the Box–Cox parameter λ. If the original observations are $Y_1, Y_2, \ldots, Y_n$, the Box–Cox transformation f_λ converts them to $f_\lambda(Y_1), f_\lambda(Y_2), \ldots, f_\lambda(Y_n)$, where

$$f_\lambda(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log(y), & \lambda = 0. \end{cases}$$

These transformations are useful when the variability of the data increases or decreases with the level. By suitable choice of λ, the variability can often be made nearly constant. In particular, for positive data whose standard deviation increases linearly with level, the variability can be stabilized by choosing $\lambda = 0$.

The choice of λ can be made visually by watching the graph of the data when you click on the pointer in the Box–Cox dialog box and drag it back and forth along the scale, which runs from zero to 1.5. Very often it is found that no transformation is needed or that the choice $\lambda = 0$ is satisfactory.

Example E.2.6. For the series AIRPASS.TSM, the variability increases with level, and the data are strictly positive. Taking natural logarithms (i.e., choosing a Box–Cox transformation with $\lambda = 0$) gives the transformed data shown in Figure E-1.

Notice how the amplitude of the fluctuations no longer increases with the level of the data. However, the seasonal effect remains, as does the upward trend. These will be removed shortly. The data stored in ITSM now consist of the natural logarithms of the original data.

□

Classical Decomposition

There are two methods provided in ITSM for the elimination of trend and seasonality. These are:

i. "classical decomposition" of the series into a trend component, a seasonal component, and a random residual component, and
ii. differencing.

Classical decomposition of the series $\{X_t\}$ is based on the model

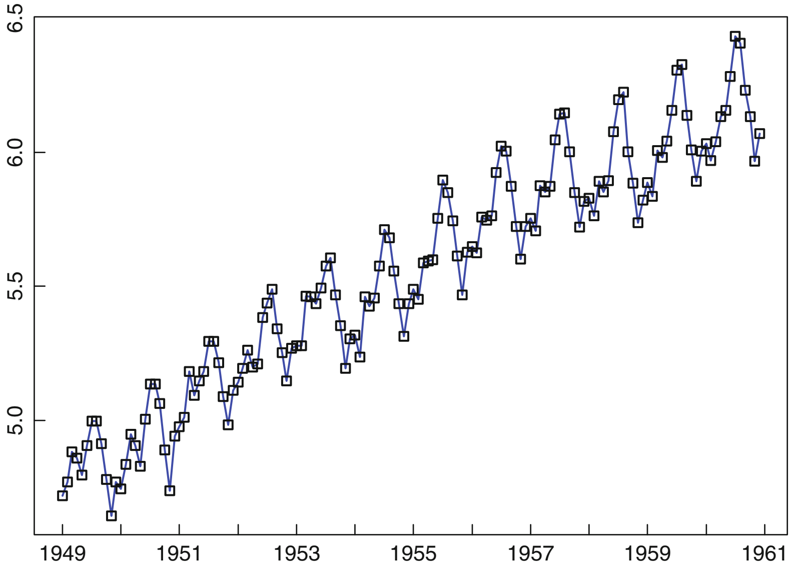

Figure E-1
The series AIRPASS.TSM after taking logs

$$X_t = m_t + s_t + Y_t,$$

where X_t is the observation at time t, m_t is a "trend component," s_t is a "seasonal component," and Y_t is a "random noise component," which is stationary with mean zero. The objective is to estimate the components m_t and s_t and subtract them from the data to generate a sequence of residuals (or estimated noise) that can then be modeled as a stationary time series.

To achieve this, select `Transform>Classical` and you will see the Classical Decomposition dialog box. To remove a seasonal component and trend, check the

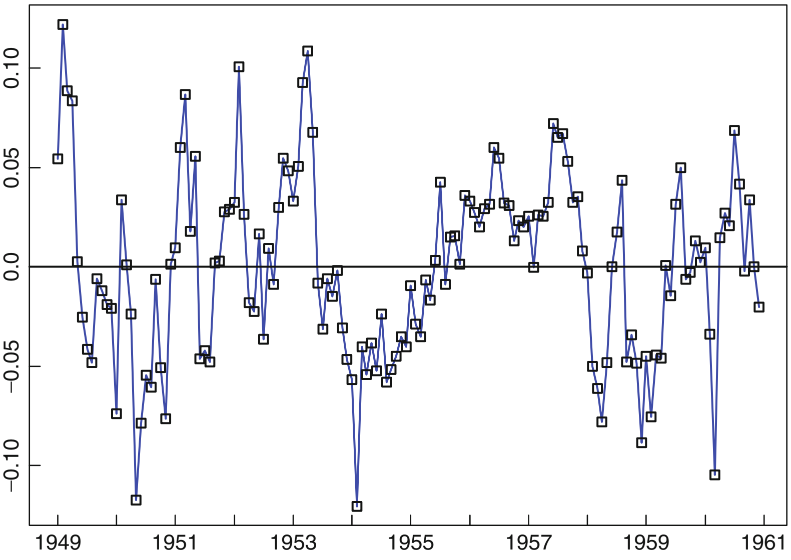

Figure E-2
The logged AIRPASS.TSM series after removal of trend and seasonal components by classical decomposition

`Seasonal Fit` and `Polynomial Fit` boxes, enter the period of the seasonal component, and choose between the alternatives `Quadratic Trend` and `Linear Trend`. Click `OK`, and the trend and seasonal components will be estimated and removed from the data, leaving the estimated noise sequence stored as the current data set.

The estimated noise sequence automatically replaces the previous data stored in ITSM.

Example E.2.7. The logged airline passenger data have an apparent seasonal component of period 12 (corresponding to the month of the year) and an approximately quadratic trend. Remove these using the option `Transform>Classical` as described above. (An alternative approach is to use the option `Regression`, which allows the specification and fitting of polynomials of degree up to 10 and a linear combination of up to 4 sine waves.)

Figure E-2 shows the transformed data (or residuals) Y_t, obtained by removal of trend and seasonality from the logged AIRPASS.TSM series by classical decomposition. $\{Y_t\}$ shows no obvious deviations from stationarity, and it would now be reasonable to attempt to fit a stationary time series model to this series. To see how well the estimated seasonal and trend components fit the data, select `Transform>Show Classical Fit`. We shall not pursue this approach any further here, but turn instead to the **differencing** approach. (You should have no difficulty in later returning to this point and completing the classical decomposition analysis by fitting a stationary time series model to $\{Y_t\}$.)

□

Differencing

Differencing is a technique that can also be used to remove seasonal components and trends. The idea is simply to consider the differences between pairs of observations with appropriate time separations. For example, to remove a seasonal component of period 12 from the series $\{X_t\}$, we generate the transformed series

$$Y_t = X_t - X_{t-12}.$$

It is clear that all seasonal components of period 12 are eliminated by this transformation, which is called **differencing at lag 12**. A linear trend can be eliminated by differencing at lag 1, and a quadratic trend by differencing twice at lag 1 (i.e., differencing once to get a new series, then differencing the new series to get a second new series). Higher-order polynomials can be eliminated analogously. It is worth noting that differencing at lag 12 eliminates not only seasonal components with period 12 but also any linear trend.

Data are differenced in ITSM by selecting `Transform>Difference` and entering the required lag in the resulting dialog box.

Example E.2.8. Restore the original airline passenger data using the option `File>Import File` and selecting AIRPASS.TSM. We take natural logarithms as in Example E.2.6 by selecting `Transform>Box-Cox` and setting $\lambda = 0$. The transformed series can now be deseasonalized by differencing at lag 12. To do this select `Transform>Difference`, enter the lag 12 in the dialog box, and click `OK`. Inspection of the graph of the deseasonalized series suggests a further differencing at lag 1 to eliminate the remaining trend. To do this, repeat the previous step with lag equal to 1 and you will see the transformed and twice-differenced series shown in Figure E-3.

□

Subtracting the Mean

The term *ARMA model* is used in ITSM to denote a *zero-mean* ARMA process (see Definition 3.1.1). To fit such a model to data, the sample mean of the data should therefore be small. Once the apparent deviations from stationarity of the data have been removed, we therefore (in most cases) subtract the sample mean of the transformed data from each observation to generate a series to which we then fit a zero-mean stationary model. Effectively we are estimating the mean of the model by the sample mean, then fitting a (zero-mean) ARMA model to the "mean-corrected" transformed data. If we know a priori that the observations are from a process with zero mean, then this process of mean correction is omitted. ITSM keeps track of all the transformations

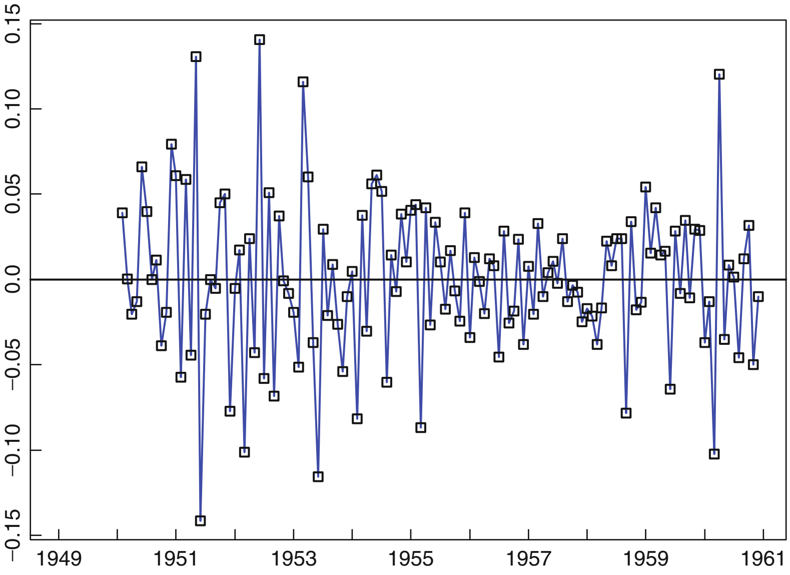

Figure E-3 The series AIRPASS.TSM after taking logs and differencing at lags 12 and 1 (including mean correction) that are made. When it comes time to predict the original series, ITSM will invert all these transformations automatically.

Example E.2.9. Subtract the mean of the transformed and twice-differenced series AIRPASS.TSM by selecting `Transform>Subtract Mean`. To check the current model status press the red INFO button, and you will see that the current model is white noise with variance 1, since no model has yet been entered.

□

E.3 Finding a Model for Your Data

After transforming the data (if necessary) as described above, we are now in a position to fit an ARMA model. ITSM uses a variety of tools to guide us in the search for an appropriate model. These include the sample ACF (autocorrelation function), the sample PACF (partial autocorrelation function), and the AICC statistic, a bias-corrected form of Akaike's AIC statistic (see Section 5.5.2).

E.3.1 Autofit

Before discussing the considerations that go into the selection, fitting, and checking of a stationary time series model, we first briefly describe an automatic feature of ITSM that searches through ARMA(p, q) models with p and q between specified limits (less than or equal to 27) and returns the model with smallest AICC value (see Sections 5.5.2 and E.3.5). Once the data set is judged to be representable by a stationary model, select `Model>Estimation>Autofit`. A dialog box will appear in which you must specify the upper and lower limits for p and q. Since the number of maximum likelihood models to be fitted is the product of the number of p-values and the number of q-values, these ranges should not be chosen to be larger than necessary. Once the limits have been specified, press `Start`, and the search will begin. You can watch the progress of the search in the dialog box that continually updates the values of p and q and the best model found so far. This option does not consider models in which the coefficients are required to satisfy constraints (other than causality) and consequently does not always lead to the optimal representation of the data. However, like the tools described below, it provides valuable information on which to base the selection of an appropriate model.

E.3.2 The Sample ACF and PACF

Pressing the second yellow button at the top of the ITSM window will produce graphs of the sample ACF and PACF for values of the lag h from 1 up to 40. For higher lags choose `Statistics>ACF/PACF>Specify Lag`, enter the maximum lag required, and click `OK`. Pressing the second yellow button repeatedly then rotates the display through ACF, PACF, and side-by-side graphs of both. Values of the ACF that decay rapidly as h increases indicate short-term dependency in the time series, while slowly decaying values indicate long-term dependency. For ARMA fitting it is desirable to have a sample ACF that decays fairly rapidly. A sample ACF that is positive and very slowly decaying suggests that the data may have a trend. A sample ACF with very slowly damped periodicity suggests the presence of a periodic seasonal component. In either of these two cases you may need to transform your data before continuing.

As a rule of thumb, the sample ACF and PACF are good estimates of the ACF and PACF of a stationary process for lags up to about a third of the sample size. It is clear from the definition of the sample ACF, $\hat{\rho}(h)$, that it will be a very poor estimator of $\rho(h)$ for h close to the sample size n.

The horizontal lines on the graphs of the sample ACF and PACF are the bounds $\pm 1.96/\sqrt{n}$. If the data constitute a large sample from an independent white noise sequence, approximately 95 % of the sample autocorrelations should lie between these bounds. Large or frequent excursions from the bounds suggest that we need a model to explain the dependence and sometimes to suggest the kind of model we need (see below). To obtain numerical values of the sample ACF and PACF, right-click on the graphs and select `Info`.

The graphs of the sample ACF and PACF sometimes suggest an appropriate ARMA model for the data. As a rough guide, if the sample ACF falls between the plotted bounds $\pm 1.96/\sqrt{n}$ for lags $h > q$, then an MA(q) model is suggested, while if the sample PACF falls between the plotted bounds $\pm 1.96/\sqrt{n}$ for lags $h > p$, then an AR(p) model is suggested.

If neither the sample ACF nor PACF "cuts off" as in the previous paragraph, a more refined model selection technique is required (see the discussion of the AICC statistic in Section 5.5.2). Even if the sample ACF or PACF does cut off at some lag, it is still advisable to explore models other than those suggested by the sample ACF and PACF values.

Example E.3.1. Figure E-4 shows the sample ACF of the AIRPASS.TSM series after taking logarithms, differencing at lags 12 and 1, and subtracting the mean. Figure E-5 shows the corresponding sample PACF. These graphs suggest that we consider an MA model of order 12 (or perhaps 23) with a large number of zero coefficients, or alternatively an AR model of order 12.

□

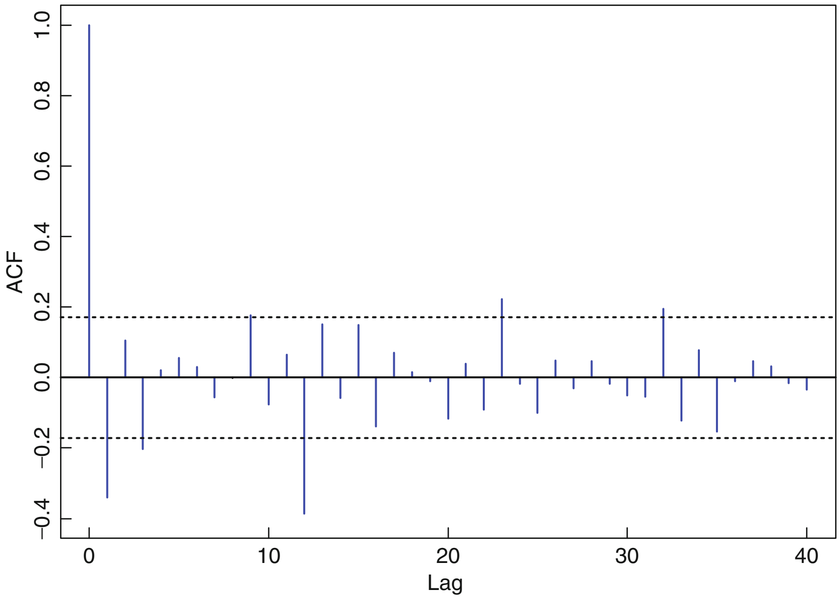

Figure E-4 The sample ACF of the transformed AIRPASS. TSM series

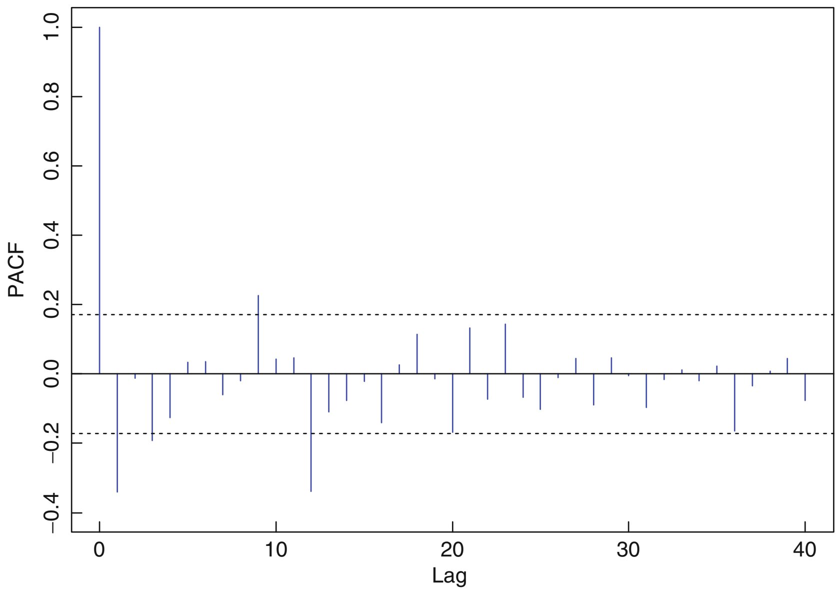

Figure E-5 The sample PACF of the transformed AIRPASS. TSM series

E.3.3 Entering a Model

A major function of ITSM is to find an ARMA model whose properties reflect to a high degree those of an observed (and possibly transformed) time series. Any particular causal ARMA(p, q) model with $p \leq 27$ and $q \leq 27$ can be entered directly by choosing `Model>Specify`, entering the values of p, q, the coefficients, and the white noise variance, and clicking `OK`. If there is a data set already open in ITSM, a quick way of entering a reasonably appropriate model is to use the option `Model>Estimation>Preliminary`, which estimates the coefficients and white noise variance of an ARMA model after you have specified the orders p and q and selected one of the four preliminary estimation algorithms available. An optimal preliminary AR model can also be fitted by checking `Find AR model with min AICC` in the `Preliminary Estimation` dialog box. If no model is entered or estimated, ITSM assumes the default ARMA(0,0), or white noise, model

$$X_t = Z_t,$$

where $\{Z_t\}$ is an uncorrelated sequence of random variables with mean zero and variance 1.

If you have data and no particular ARMA model in mind, it is advisable to use the option `Model>Estimation>Preliminary` or equivalently to press the blue PRE button at the top of the ITSM window.

Sometimes you may wish to try a model found in a previous session or a model suggested by someone else. In that case choose `Model>Specify` and enter the required model. You can save both the model and data from any project by selecting `File>Project>Save as` and specifying the name for the new file. When the new file is opened, both the model and the data will be imported. To create a project with this model and a new data set select `File>Import File` and enter the name of the file containing the new data. (This file must contain data only. If it also contains a model, then the model will be imported with the data and the model previously in ITSM will be overwritten.)

E.3.4 Preliminary Estimation

The option `Model>Estimation>Preliminary` contains fast (but not the most efficient) model-fitting algorithms. They are useful for suggesting the most promising models for the data, but should be followed by maximum likelihood estimation using `Model>Estimation>Max likelihood`. The fitted preliminary model is generally used as an initial approximation with which to start the nonlinear optimization carried out in the course of maximizing the (Gaussian) likelihood.

To fit an ARMA model of specified order, first enter the values of p and q (see Section 2.6.1). For pure AR models $q = 0$, and the preliminary estimation option offers a choice between the Burg and Yule–Walker estimates. (The Burg estimates frequently give higher values of the Gaussian likelihood than the Yule–Walker estimates.) If $q = 0$, you can also check the box `Find AR model with min AICC` to allow the program to fit AR models of orders $0, 1, \ldots, 27$ and select the one with smallest AICC value (Section 5.5.2). For models with $q > 0$, ITSM provides a choice between two preliminary estimation methods, one based on the Hannan–Rissanen procedure and the other on the innovations algorithm. If you choose the innovations option, a default value of m will be displayed on the screen. This parameter was defined in Section 5.1.3. The standard choice is the default value computed by ITSM. The Hannan–Rissanen algorithm is recommended when p and q are both greater than 0, since it tends to give causal models more frequently than the innovations method. The latter is recommended when $p = 0$.

Once the required entries in the Preliminary Estimation dialog box have been completed, click `OK`, and ITSM will quickly estimate the parameters of the selected model and display a number of diagnostic statistics. (If p and q are both greater than 0, it is possible that the fitted model may be noncausal, in which case ITSM sets all the coefficients to .001 to ensure the causality required for subsequent maximum likelihood estimation. It will also give you the option of fitting a model of different order.)

Provided that the fitted model is causal, the estimated parameters are given with the ratio of each estimate to 1.96 times its standard error. The denominator (1.96 × standard error) is the critical value (at level .05) for the coefficient. Thus, if the ratio is greater than 1 in absolute value, we may conclude (at level .05) that the corresponding coefficient is different from zero. On the other hand, a ratio less than 1 in absolute value suggests the possibility that the corresponding coefficient in the model may be zero. (If the innovations option is chosen, the ratios of estimates to 1.96 × standard error are displayed only when $p = q$ or $p = 0$.) In the Preliminary Estimates window you will also see one or more estimates of the white noise variance (the residual sum of squares divided by the sample size is the estimate retained by ITSM) and some further diagnostic statistics. These are $-2\ln L\big(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2\big)$, where L denotes the Gaussian likelihood (5.2.9), and the AICC statistic

$$-2\ln L + 2(p + q + 1)n/(n - p - q - 2)$$

(see Section 5.5.2).

Our eventual aim is to find a model with as small an AICC value as possible. Smallness of the AICC value computed in the preliminary estimation phase is indicative of a good model, but should be used only as a rough guide. Final decisions between models should be based on maximum likelihood estimation, carried out using the option `Model>Estimation>Max likelihood`, since for fixed p and q, the values of $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, and σ^2 that minimize the AICC statistic are the maximum likelihood estimates, not the preliminary estimates. After completing preliminary estimation, ITSM stores the estimated model coefficients and white noise variance. The stored estimate of the white noise variance is the sum of squares of the residuals (or one-step prediction errors) divided by the number of observations.

A variety of models should be explored using the preliminary estimation algorithms, with a view to finding the most likely candidates for minimizing AICC when the parameters are reestimated by maximum likelihood.

Example E.3.2. To find the minimum-AICC Burg AR model for the logged, differenced, and mean-corrected series AIRPASS.TSM currently stored in ITSM, press the blue PRE button, set the MA order equal to zero, select `Burg` and `Find AR model with min AICC`, and then click `OK`. The minimum-AICC AR model is of order 12 with an AICC value of −458.13. To fit a preliminary MA(25) model to the same data, press the blue PRE button again, but this time set the AR order to 0, the MA order to 25, select `Innovations`, and click `OK`.

The ratios (estimated coefficient)/(1.96× standard error) indicate that the coefficients at lags 1 and 12 are nonzero, as suggested by the sample ACF. The estimated coefficients at lags 3 and 23 also look substantial even though the corresponding ratios are less than 1 in absolute value. The displayed values are as follows:

```
MA COEFFICIENTS
     −0.3568      0.0673     −0.1629     −0.0415      0.1268
      0.0264      0.0283     −0.0648      0.1326     −0.0762
     −0.0066     −0.4987      0.1789     −0.0318      0.1476
     −0.1461      0.0440     −0.0226     −0.0749     −0.0456
     −0.0204     −0.0085      0.2014     −0.0767     −0.0789
RATIO OF COEFFICIENTS TO (1.96*STANDARD ERROR)
     −2.0833      0.3703     −0.8941     −0.2251      0.6875
      0.1423      0.1522     −0.3487      0.7124     −0.4061
     −0.0353     −2.6529      0.8623     −0.1522      0.7068
     −0.6944      0.2076     −0.1065     −0.3532     −0.2147
     −0.0960     −0.0402      0.9475     −0.3563     −0.3659
```

The estimated white noise variance is 0.00115 and the AICC value is −440.93, which is not as good as that of the AR(12) model. Later we shall find a subset MA(25) model that has a smaller AICC value than both of these models.

□

E.3.5 The AICC Statistic

The AICC statistic for the model with parameters p, q, ϕ, and $\boldsymbol{\theta}$ is defined (see Section 5.5.2) as

$$\mathrm{AICC}(\phi, \boldsymbol{\theta}) = -2 \ln L(\phi, \boldsymbol{\theta}, S(\phi, \boldsymbol{\theta})/n) + 2(p+q+1)n/(n-p-q-2),$$

and a model chosen according to the AICC criterion minimizes this statistic.

Model-selection statistics other than AICC are also available in ITSM. A Bayesian modification of the AIC statistic known as the BIC statistic is also computed in the option `Model>Estimation>Max likelihood`. It is used in the same way as the AICC.

An exhaustive search for a model with minimum AICC or BIC value can be very slow. For this reason the sample ACF and PACF and the preliminary estimation techniques described above are useful in narrowing down the range of models to be considered more carefully in the maximum likelihood estimation stage of model fitting.

E.3.6 Changing Your Model

The model currently stored by the program can be checked at any time by selecting `Model>Specify`. Any parameter can be changed in the resulting dialog box, including the white noise variance. The model can be filed together with the data for later use by selecting `File>Project>Save as` and specifying a file name with suffix .TSM.

Example E.3.3. We shall now set some of the coefficients in the current model to zero. To do this choose `Model>Specify` and click on the box containing the value −0.35676 of Theta(1). Press `Enter`, and the value of Theta(2) will appear in the box. Set this to zero. Press `Enter` again, and the value of Theta(3) will appear. Continue to work through the coefficients, setting all except Theta(1), Theta(3), Theta(12), and Theta(23) equal to zero. When you have reset the parameters, click `OK`, and the new model stored in ITSM will be the subset MA(23) model

$$X_t = Z_t - 0.357Z_{t-1} - 0.163Z_{t-3} - 0.499Z_{t-12} + 0.201Z_{t-23},$$

where $\{Z_t\} \sim \text{WN}(0, 0.00115)$.

□

E.3.7 Maximum Likelihood Estimation

Once you have specified values of p and q and possibly set some coefficients to zero, you can carry out efficient parameter estimation by selecting `Model>Estimation>Max likelihood` or equivalently by pressing the blue MLE button.

The resulting dialog box displays the default settings, which in most cases will not need to be modified. However, if you wish to compute the likelihood without maximizing it, check the box labeled `No optimization`. The remaining information concerns the optimization settings. (With the default settings, any coefficients that are set to zero will be treated as fixed values and not as parameters. Coefficients to be optimized must therefore not be set exactly to zero. If you wish to impose further constraints on the optimization, press the `Constrain optimization` button. This allows you to fix certain coefficients or to impose multiplicative relationships on the coefficients during optimization.)

To find the maximum likelihood estimates of your parameters, click `OK`, and the estimated parameters will be displayed. To refine the estimates, repeat the estimation, specifying a smaller value of the accuracy parameter in the `Maximum Likelihood` dialog box.

Example E.3.4. To find the maximum likelihood estimates of the parameters in the model for the logged, differenced, and mean-corrected airline passenger data currently stored in ITSM, press the blue MLE button and click `OK`. The following estimated parameters and diagnostic statistics will then be displayed:

```
ARMA MODEL:
X(t) = Z(t)+(−.355) ∗ Z(t − 1)+(−.201) ∗ Z(t − 3)+(−.523) ∗ Z(t − 12)
        +(.242) ∗ Z(t − 23)
WN Variance = .001250

MA Coefficients
THETA( 1)= -.355078 THETA( 3)= -.201125
THETA(12)= -.523423 THETA(23)= .241527
Standard Error of MA Coefficients
THETA( 1): .059385 THETA( 3): .059297
THETA(12): .058011 THETA(23): .055828

(Residual SS)/N =   .125024E−02
AICC = -.486037E+03
BIC = -.487622E+03

-2 Ln(Likelihood)= -.496517E+03

Accuracy parameter = .00205000

Number of iterations = 5

Number of function evaluations = 46

Optimization stopped within accuracy level.
```

The last message indicates that the minimum of $-2\ln L$ has been located with the specified accuracy. If you see the message

```
Iteration limit exceeded,
```

then the minimum of $-2\ln L$ could not be located with the number of iterations (50) allowed. You can continue the search (starting from the point at which the iterations were interrupted) by pressing the MLE button to continue the minimization and possibly increasing the maximum number of iterations from 50 to 100.

□

E.3.8 Optimization Results

After maximizing the Gaussian likelihood, ITSM displays the model parameters (coefficients and white noise variance), the values of $-2\ln L$, AICC, BIC, and information regarding the computations.

Example E.3.5. The next stage of the analysis is to consider a variety of competing models and to select the most suitable. The following table shows the AICC statistics for a variety of subset moving average models of order less than 24.

Lags						AICC
1	3		12		23	−486.04
1	3		12	13	23	−485.78
1	3	5	12		23	−489.95
1	3		12	13		−482.62
1			12			−475.91

The best of these models from the point of view of AICC value is the one with nonzero coefficients at lags 1, 3, 5, 12, and 23. To obtain this model from the one currently stored in ITSM, select `Model>Specify`, change the value of THETA(5) from zero to .001, and click `OK`. Then reoptimize by pressing the blue MLE button and clicking `OK`. You should obtain the noninvertible model

$$X_t = Z_t - 0.434Z_{t-1} - 0.305Z_{t-3} + 0.238Z_{t-5} - 0.656Z_{t-12} + 0.351Z_{t-23},$$

where $\{Z_t\} \sim \text{WN}(0, 0.00103)$. For future reference, file the model and data as AIRPASS2.TSM using the option `File>Project>Save as`.

□

The next step is to check our model for goodness of fit.

E.4 Testing Your Model

Once we have a model, it is important to check whether it is any good or not. Typically this is judged by comparing observations with corresponding predicted values obtained from the fitted model. If the fitted model is appropriate then the prediction errors should behave in a manner that is consistent with the model. The **residuals** are the rescaled one-step prediction errors,

$$\hat{W}_t = (X_t - \hat{X}_t)/\sqrt{r_{t-1}},$$

where $\hat{X}_t$ is the best linear mean-square predictor of X_t based on the observations up to time $t-1$, $r_{t-1} = E(X_t - \hat{X}_t)^2/\sigma^2$ and σ^2 is the white noise variance of the fitted model.

If the data were truly generated by the fitted ARMA(p, q) model with white noise sequence $\{Z_t\}$, then for large samples the properties of $\{\hat{W}_t\}$ should reflect those of $\{Z_t\}$. To check the appropriateness of the model we therefore examine the residual series $\{\hat{W}_t\}$, and check that it resembles a realization of a white noise sequence.

ITSM provides a number of tests for doing this in the Residuals Menu, which is obtained by selecting the option `Statistics>Residual Analysis`. Within this option are the suboptions

Plot
QQ-Plot (normal)
QQ-Plot (t-distr)
Histogram
ACF/PACF
ACF Abs vals/Squares
Tests of randomness

E.4.1 Plotting the Residuals

Select `Statistics>Residual Analysis>Histogram`, and you will see a histogram of the **rescaled residuals**, defined as

$$\hat{R}_t = \hat{W}_t/\hat{\sigma},$$

where $n\hat{\sigma}^2$ is the sum of the squared residuals. If the fitted model is appropriate, the histogram of the rescaled residuals should have mean close to zero. If in addition the data are Gaussian, this will be reflected in the shape of the histogram, which should then resemble a normal density with mean zero and variance 1.

Select `Statistics>Residual Analysis>Plot` and you will see a graph of $\hat{R}_t$ vs. t. If the fitted model is appropriate, this should resemble a realization of a white noise sequence. Look for trends, cycles, and nonconstant variance, any of which suggest that the fitted model is inappropriate. If substantially more than 5 % of

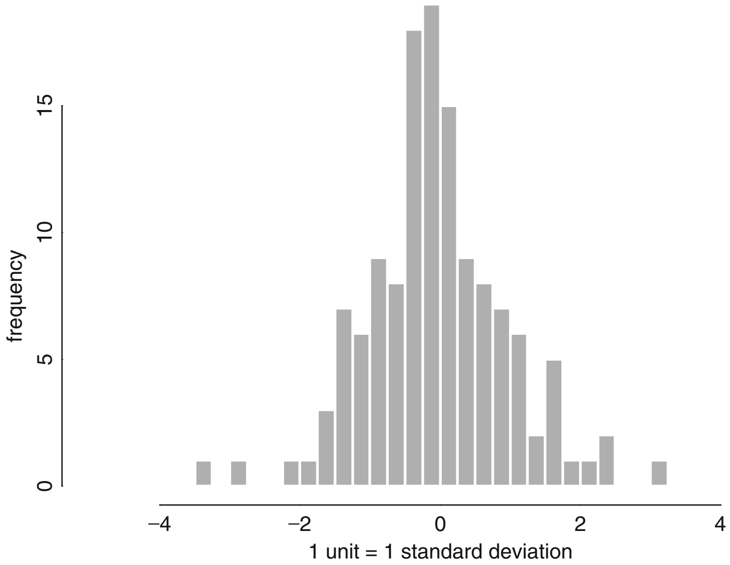

Figure E-6
Histogram of the rescaled residuals from AIRPASS.MOD

the rescaled residuals lie outside the bounds ±1.96 or if there are rescaled residuals far outside these bounds, then the fitted model should not be regarded as Gaussian.

Compatibility of the distribution of the residuals with either the normal distribution or the t-distribution can be checked by inspecting the corresponding qq plots and checking for approximate linearity. To test for normality, the Jarque–Bera statistic is also computed.

Example E.4.1. The histogram of the rescaled residuals from our model for the logged, differenced, and mean-corrected airline passenger series is shown in Figure E-6. The mean is close to zero, and the shape suggests that the assumption of Gaussian white noise is not unreasonable in our proposed model.

The graph of $\hat{R}_t$ vs. t is shown in Figure E-7. A few of the rescaled residuals are greater in magnitude than 1.96 (as is to be expected), but there are no obvious indications here that the model is inappropriate. The approximate linearity of the normal qq plot and the Jarque–Bera test confirm the approximate normality of the residuals.

□

E.4.2 ACF/PACF of the Residuals

If we were to assume that our fitted model is the true process generating the data, then the observed residuals would be realized values of a white noise sequence.

In particular, the sample ACF and PACF of the observed residuals should lie within the bounds $\pm 1.96/\sqrt{n}$ roughly 95 % of the time. These bounds are displayed on the graphs of the ACF and PACF. If substantially more than 5 % of the correlations are outside these limits, or if there are a few very large values, then we should look for a better-fitting model. (More precise bounds, due to Box and Pierce, can be found in Brockwell and Davis (1991) Section 10.4.)

Example E.4.2. Choose `Statistics>Residual Analysis>ACF/PACF`, or equivalently press the middle green button at the top of the ITSM window. The sample ACF and PACF of the residuals will then appear as shown in Figures E-8 and E-9. No correlations

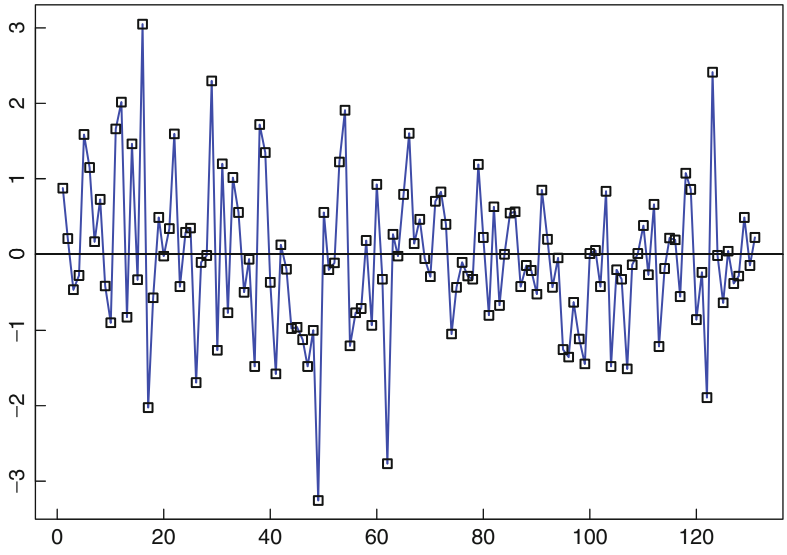

Figure E-7
Time plot of the rescaled residuals from AIRPASS.MOD

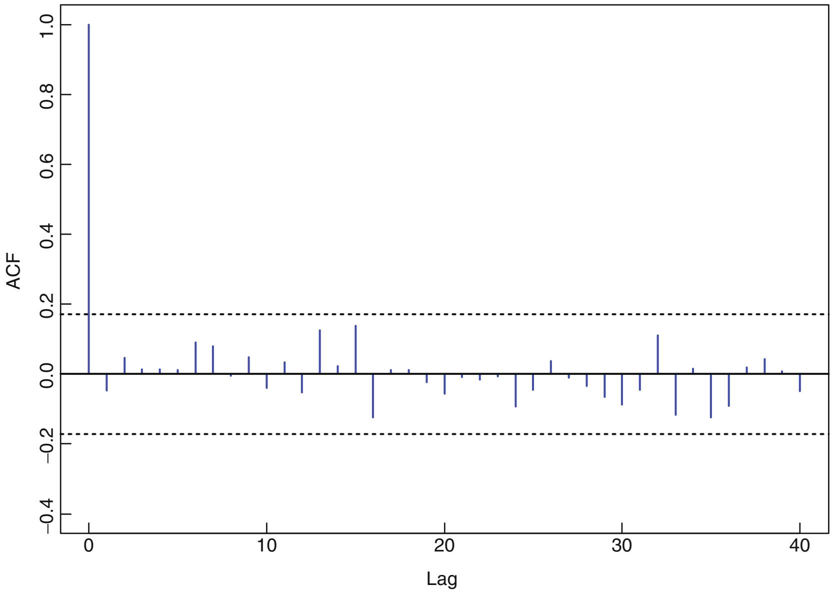

Figure E-8
Sample ACF of the residuals from AIRPASS.MOD

are outside the bounds in this case. They appear to be compatible with the hypothesis that the residuals are in fact observations of a white noise sequence. To check for independence of the residuals, the sample autocorrelation functions of their absolute values and squares can be plotted by clicking on the third green button.

□

E.4.3 Testing for Randomness of the Residuals

The option `Statistics>Residual Analysis>Tests of Randomness` carries out the six tests for randomness of the residuals described in Section 5.3.3.

Example E.4.3. The residuals from our model for the logged, differenced, and mean-corrected series AIRPASS.TSM are checked by selecting the option indicated above and selecting the parameter h for the portmanteau tests. Adopting the value $h = 25$ suggested by ITSM, we obtain the following results:

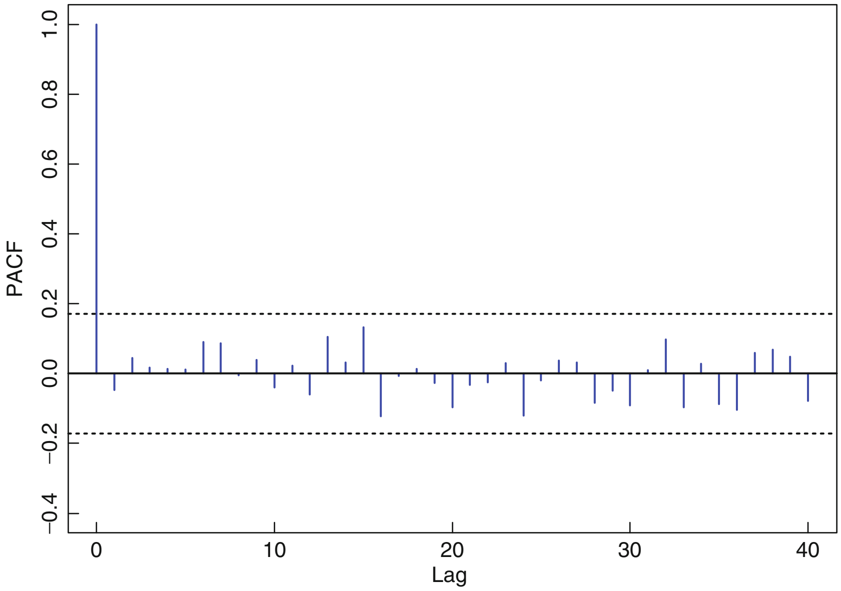

Figure E-9 Sample PACF of the residuals from AIRPASS.MOD

```
RANDOMNESS TEST STATISTICS (see Section 5.3.3)
_______________________________________________

LJUNG-BOX PORTM.= 13.76      CHISQUR( 20),                  p-value = 0.843
MCLEOD-LI PORTM.= 17.39      CHISQUR( 25),                  p-value = 0.867
TURNING POINTS = 87.         ANORMAL( 86.00, 4.79**2),      p-value = 0.835
DIFFERENCE-SIGN = 65.        ANORMAL( 65.00, 3.32**2),      p-value = 1.000
RANK TEST = 3934.            ANORMAL(4257.50, 251.3**2),    p-value = 0.198
JARQUE–BERA = 4.33           CHISQUR(2)                     p-value = 0.115
ORDER OF MIN AICC            YW MODEL FOR RESIDUALS = 0
```

Every test is easily passed by our fitted model (with significance level $\alpha = 0.05$), and the order of the minimum-AICC AR model for the residuals supports the compatibility of the residuals with white noise. For later use, file the residuals by pressing the red EXP button and exporting the residuals to a file with the name AIRRES.TSM.

□

E.5 Prediction

One of the main purposes of time series modeling is the prediction of future observations. Once you have found a suitable model for your data, you can predict future values using the option `Forecasting>ARMA`. (The other options listed under `Forecasting` refer to the methods of Chapter 10.)

E.5.1 Forecast Criteria

Given observations $X_1, \ldots, X_n$ of a series that we assume to be appropriately modeled as an ARMA(p, q) process, ITSM predicts future values of the series X_{n+h} from the data and the model by computing the linear combination $P_n(X_{n+h})$ of $X_1, \ldots, X_n$ that minimizes the mean squared error $E(X_{n+h} - P_n(X_{n+h}))^2$.

E.5.2 Forecast Results

Assuming that the current data set has been adequately fitted by the current ARMA(p, q) model, choose `Forecasting>ARMA`, and you will see the `ARMA Forecast` dialog box.

You will be asked for the number of forecasts required, which of the transformations you wish to invert (the default settings are to invert all of them so as to obtain forecasts of the *original* data), whether or not you wish to plot prediction bounds (assuming normality), and if so, the confidence level required, e.g., 95 %. After providing this information, click `OK`, and the data will be plotted with the forecasts (and possibly prediction bounds) appended. As is to be expected, the separation of the prediction bounds increases with the lead time h of the forecast.

Right-click on the graph, select `Info`, and the numerical values of the predictors and prediction bounds will be printed.

Example E.5.1. We left our logged, differenced, and mean-corrected airline passenger data stored in ITSM with the subset MA(23) model found in Example D.3.5. To predict the next 24 values of the original series, select `Forecasting>ARMA` and accept the default settings in the dialog box by clicking `OK`. You will then see the graph shown in Figure E-10. Numerical values of the forecasts are obtained by right-clicking on the graph and selecting `Info`. The ARMA `Forecast` dialog box also permits using a model constructed from a subset of the data to obtain forecasts and prediction bounds for the remaining observed values of the series.

□

E.6 Model Properties

ITSM can be used to analyze the properties of a specified ARMA process without reference to any data set. This enables us to explore and compare the properties of different ARMA models in order to gain insight into which models might best represent particular features of a given data set.

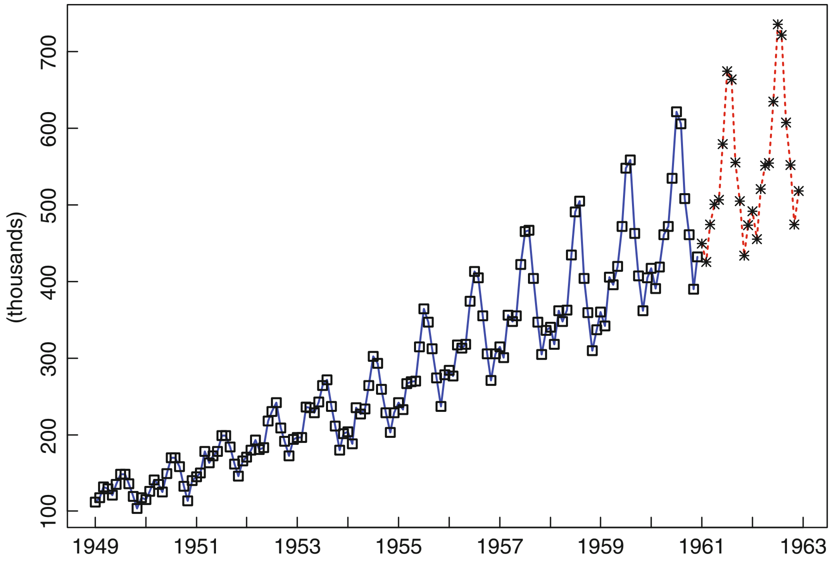

Figure E-10 The original AIRPASS data with 24 forecasts appended

For any ARMA(p, q) process or fractionally integrated ARMA(p, q) process with $p \leq 27$ and $q \leq 27$, ITSM allows you to compute the autocorrelation and partial autocorrelation functions, the spectral density and distribution functions, and the MA(∞) and AR(∞) representations of the process. It also allows you to generate simulated realizations of the process driven by either Gaussian or non-Gaussian noise. The use of these options is described in this section.

Example E.6.1. We shall illustrate the use of ITSM for model analysis using the model for the transformed series AIRPASS.TSM that is currently stored in the program. □

E.6.1 ARMA Models

For modeling zero-mean stationary time series, ITSM uses the class of ARMA (and fractionally integrated ARMA) processes. ITSM enables you to compute characteristics of the causal ARMA model defined by

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q},$$

or more concisely $\phi(B)X_t = \theta(B)Z_t$, where $\{Z_t\} \sim \mathrm{WN}\left(0, \sigma^2\right)$ and the parameters are all specified. (Characteristics of the fractionally integrated ARIMA(p, d, q) process defined by

$$(1 - B)^d \phi(B)X_t = \theta(B)Z_t, \quad |d| < 0.5,$$

can also be computed.)

ITSM works exclusively with causal models. It will not permit you to enter a model for which $1-\phi_1 z-\cdots-\phi_p z^p$ has a zero inside or on the unit circle, nor does it generate fitted models with this property. From the point of view of second-order properties, this represents no loss of generality (Section 3.1). If you are trying to enter an ARMA(p, q) model manually, the simplest way to ensure that your model is causal is to set all the autoregressive coefficients close to zero (e.g., .001). ITSM will not accept a noncausal model.

ITSM does not restrict models to be invertible. You can check whether or not the current model is invertible by choosing `Model>Specify` and pressing the button labeled Causal/Invertible in the resulting dialog box. If the model is noninvertible, i.e., if the moving-average polynomial $1 + \theta_1 z + \cdots + \theta_q z^q$ has a zero inside or on the unit circle, the message `Non-invertible` will appear beneath the box containing the moving-average coefficients. (A noninvertible model can be converted to an invertible model with the same autocovariance function by choosing `Model>Switch to invertible`. If the model is already invertible, the program will tell you.)

E.6.2 Model ACF, PACF

The *model* ACF and PACF are plotted using `Model>ACF/PACF>Model`. If you wish to change the maximum lag from the default value of 40, select `Model>ACF/PACF>Specify Lag` and enter the required maximum lag. (It can be much larger than 40, e.g., 10,000). The graph will then be modified, showing the correlations up to the specified maximum lag.

If there is a data file open as well as a model in ITSM, the model ACF and PACF can be compared with the sample ACF and PACF by pressing the third yellow button at the top of the ITSM window. The model correlations will then be plotted in red, with the corresponding sample correlations shown in the same graph but plotted in green.

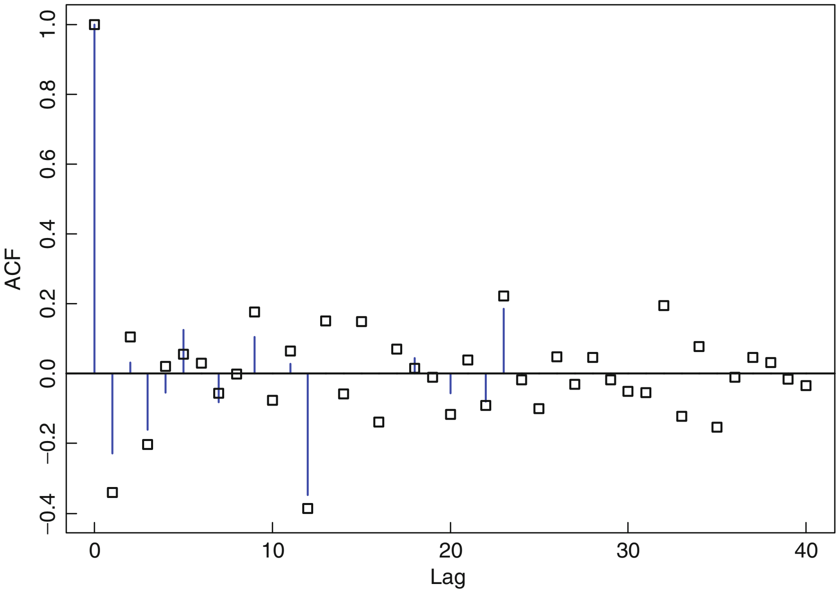

Figure E-11
The ACF of the model in Example E.3.5 together with the sample ACF of the transformed AIRPASS.TSM series

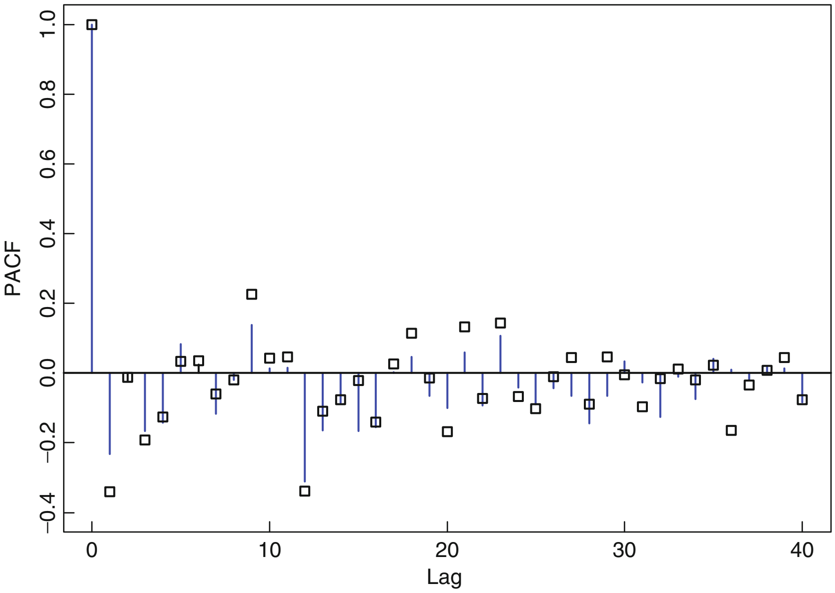

Figure E-12
The PACF of the model in Example E.3.5 together with the sample PACF of the transformed AIRPASS.TSM series

Example E.6.2. The sample and model ACF and PACF for the current model and transformed series AIRPASS.TSM are shown in Figures E-11 and E-12. They are obtained by pressing the third yellow button at the top of the ITSM window. The vertical lines represent the model values, and the squares are the sample ACF/PACF. The graphs show that the data and the model ACF both have large values at lag 12, while the sample and model partial autocorrelation functions both tend to die away geometrically after the peak at lag 12. The similarities between the graphs indicate that the model is capturing some of the important features of the data.

□

E.6.3 Model Representations

As indicated in Section 3.1, if $\{X_t\}$ is a causal ARMA process, then it has an MA(∞) representation

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}, \quad t = 0, \pm 1, \pm 2, \ldots,$$

where $\sum\limits_{j=0}^{\infty} |\psi_j| < \infty$ and $\psi_0 = 1$.

Similarly, if $\{X_t\}$ is an invertible ARMA process, then it has an AR(∞) representation

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}, \quad t = 0, \pm 1, \pm 2, \ldots,$$

where $\sum_{j=0}^{\infty} |\pi_j| < \infty$ and $\pi_0 = 1$.

For any specified causal ARMA model you can determine the coefficients in these representations by selecting the option `Model>AR/MA Infinity`. (If the model is not invertible, you will see only the MA(∞) coefficients, since the AR(∞) representation does not exist in this case.)

Example E.6.3. The current subset MA(23) model for the transformed series AIRPASS.TSM does not have an AR(∞) representation, since it is not invertible. However, we can replace the model with an invertible one having the same autocovariance function by selecting `Model>Switch to Invertible`. For this model we can then find an AR(∞) representation by selecting `Model>AR Infinity`. This gives 50 coefficients, the first 20 of which are shown below.

MA − Infinity	AR − Infinity	
j	*psi*(j)	*pi*(j)
0	1.00000	1.00000
1	−0.36251	0.36251
2	0.01163	0.11978
3	−0.26346	0.30267
4	−0.06924	0.27307
5	0.15484	−0.00272
6	−0.02380	0.05155
7	−0.06557	0.16727
8	−0.04487	0.10285
9	0.01921	0.01856
10	−0.00113	0.07947
11	0.01882	0.07000
12	−0.57008	0.58144
13	0.00617	0.41683
14	0.00695	0.23490
15	0.03188	0.37200
16	0.02778	0.38961
17	0.01417	0.10918
18	0.02502	0.08776
19	0.00958	0.22791

□

E.6.4 Generating Realizations of a Random Series

ITSM can be used to generate realizations of a random time series defined by the currently stored model.

To generate such a realization, select the option `Model>Simulate`, and you will see the ARMA Simulation dialog box. You will be asked to specify the number of observations required, the white noise variance (if you wish to change it from the current value), and an integer-valued random number seed (by specifying and recording this integer with up to nine digits you can reproduce the same realization at a later date by reentering the same seed). You will also have the opportunity to add a specified mean to the simulated ARMA values. If the current model has been fitted to transformed data, then you can also choose to apply the inverse transformations to the simulated ARMA to generate a simulated version of the *original* series. The default distribution for the white noise is Gaussian. However, by pressing the button `Change noise distribution` you can select from a variety of alternative distributions or by checking the box `Use Garch model for noise process` you can generate an ARMA process driven by GARCH noise. Finally, you can choose whether the simulated data will overwrite the data set in the current project or whether they will be used to create a new project. Once you are satisfied with your choices, click `OK`, and the simulated series will be generated.

Example E.6.4. To generate a simulated realization of the series AIRPASS.TSM using the current model and transformed data set, select the option `Model>Simulate`. The default options in the dialog box are such as to generate a realization of the *original* series as a new project, so it suffices to click `OK`. You will then see a graph of the simulated series that should resemble the original series AIRPASS.TSM.

□

E.6.5 Spectral Properties

Spectral properties of both data and fitted ARMA models can also be computed and plotted with the aid of ITSM. The spectral density of the *model* is determined by selecting the option `Spectrum>Model`. Estimation of the spectral density from observations of a stationary series can be carried out in two ways, either by fitting an ARMA model as already described and computing the spectral density of the fitted model (Section 4.4) or by computing the periodogram of the data and smoothing (Section 4.2). The latter method is applied by selecting the option `Spectrum>Smoothed Periodogram`. Examples of both approaches are given in Chapter 4.

E.7 Multivariate Time Series

Observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of an m-component time series must be stored as an ASCII file with n rows and m columns, with at least one space between entries in the same row. To open a multivariate series for analysis, select `File>Project>Open>Multivariate` and click `OK`. Then double-click on the file containing the data, and you will be asked to enter the number of columns (m) in the data file. After doing this, click `OK`, and you will see graphs of each component of the series, with the multivariate tool bar at the top of the ITSM screen. For examples of the application of ITSM to the analysis of multivariate series, see Chapter 8.

References

Akaike, H.(1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics, 21*, 243–247.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *2nd International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Akaike, H. (1978). Time series analysis and control through parametric models. In D. F. Findley (Ed.), *Applied time series analysis*. New York: Academic.

Andersen, T. G., & Benzoni, L. (2009). Realized volatility. In T. G. Andersen, R. A. Davis, J.-P. Kreiss, & T. V. Mikosch (Eds.), *Handbook of financial time series* (pp. 555–576). Berlin, Heidelberg: Springer.

Andersen, T. G., Davis, R. A., Kreiss, J.-P., & Mikosch, T. V. (Eds.) (2009). *Handbook of financial time series*. Berlin: Springer.

Anderson, T. W. (1971). *The statistical analysis of time series*. New York: Wiley.

Anderson, T. W. (1980). Maximum likelihood estimation for vector autoregressive moving-average models. In D. R. Brillinger & G. C. Tiao (Eds.), *Directions in time series* (pp. 80–111). Beachwood: Institute of Mathematical Statistics.

Ansley, C. F. (1979). An algorithm for the exact likelihood of a mixed autoregressive-moving-average process. *Biometrika, 66*, 59–65.

Ansley, C. F., & Kohn, R. (1985). On the estimation of ARIMA models with missing values. In E. Parzen (Ed.), *Time series analysis of irregularly observed data*. Springer lecture notes in statistics (Vol. 25, pp. 9–37), Springer-Verlag, Berlin, Heidelberg, New York.

Aoki, M. (1987). *State space modeling of time series*. Berlin: Springer.

Applebaum, D. *Lévy processes and stochastic calculus*. Cambridge: Cambridge University Press.

Atkins, S. M. (1979). Case study on the use of intervention analysis applied to traffic accidents. *Journal of the Operations Research Society, 30*(7), 651–659.

Bachelier, L. (1900). Théorie de la spéculation. *Annales de lÉcole Normale Supérieure, 17*, 21–86.

Baillie, R. T., Bollerslev, T., & Mikkelsen, H. O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 74*, 3–30.

Barndorff-Nielsen, O. E. (1978). *Information and exponential families in statistical theory*. New York: Wiley.

P.J. Brockwell, R.A. Davis, *Introduction to Time Series and Forecasting*,
Springer Texts in Statistics, DOI 10.1007/978-3-319-29854-2

Barndorff-Niesen, O. E., & Shephard, N. (2001). Non-Gaussian Ornstein–Uhlenbeck based models and some of their uses in financial economics (with discussion). *Journal of the Royal Statistical Society Series B, 63*, 167–241.

Bertoin, J. (1996). *Lévy processes*. Cambridge: Cambridge University Press.

Bhattacharyya, M. N., & Layton, A. P. (1979). Effectiveness of seat belt legislation on the Queensland road toll—An Australian case study in intervention analysis. *Journal of the American Statistical Association, 74*, 596–603.

Billingsley, P. (1995). *Probability and measure* (3rd ed.). New York: Wiley.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy, 81*, 637–654.

Bloomfield, P. (2000). *Fourier analysis of time series: An introduction* (2nd ed.). New York: Wiley.

Bollerslev, T. (1986), Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*, 307–327.

Bollerslev, T., & Mikkelsen, H. O. (1996). Modeling and pricing long memory in stock market volatility. *Journal of Econometrics, 73*, 151–184.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B, 26*, 211–252.

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control (revised edition)*. San Francisco: Holden-Day.

Box, G. E. P., & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving-average time series models. *Journal of the American Statistical Association, 65*, 1509–1526.

Box, G. E. P., & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association, 70*, 70–79.

Breidt, F. J., & Davis, R. A. (1992). Time reversibility, identifiability and independence of innovations for stationary time series. *Journal of Time Series Analysis, 13*, 377–390.

Brockwell, P.J. (2014), Recent results in the theory and applications of CARMA processes, *Ann. Inst. Stat. Math. 66*, 637–685.

Brockwell, P. J., Chadraa, E., & Lindner, A. (2006). Continuous-time GARCH processes. *Annals of Applied Probability, 16*, 790–826.

Brockwell, P. J., & Davis, R. A. (1988). Applications of innovation representations in time series analysis. In J. N. Srivastava (Ed.), *Probability and statistics, essays in honor of Franklin A. Graybill* (pp. 61–84). Amsterdam: Elsevier.

Brockwell, P. J., & Davis, R. A. (1991). *Time series: Theory and methods* (2nd ed.). New York: Springer.

Brockwell, P. J., & Lindner, A. (2009). Existence and uniqueness of stationary Lévy-driven CARMA processes. *Stochastic Processes and Their Applications, 119*, 2660–2681.

Brockwell, P. J., & Lindner, A. (2012). Integration of CARMA processes and spot volatility modelling. *Journal of Time Series Analysis, 34*, 156–167.

Campbell, J., Lo, A., & McKinlay, C. (1996). *The econometrics of financial markets*. Princeton, NJ: Princeton University Press.

Chan, K. S., & Ledolter, J. (1995). Monte Carlo EM estimation for time series models involving counts. *Journal of the American Statistical Association, 90*, 242–252.

Chan, K. S., & Tong, H. (1987). A note on embedding a discrete parameter ARMA model in a continuous parameter ARMA model. *Journal of Time Series Analysis, 8*, 277–281.

Cochran, D., & Orcutt, G. H. (1949). Applications of least squares regression to relationships containing autocorrelated errors. *Journal of the American Statistical Association, 44*, 32–61.

Davis, M., & Etheridge, A. (2006). *Louis Bachelier's theory of speculation: The origins of modern finance*. Princeton, NJ: Princeton University Press.

Davis, M. H. A., & Vinter, R. B. (1985). *Stochastic modelling and control*. London: Chapman and Hall.

Davis, R. A., Chen, M., & Dunsmuir, W. T. M. (1995). Inference for MA(1) processes with a root on or near the unit circle. *Probability and Mathematical Statistics, 15*, 227–242.

Davis, R. A., Chen, M., & Dunsmuir, W. T. M. (1996). Inference for seasonal moving-average models with a unit root. In *Athens conference on applied probability and time series, volume 2: Time series analysis*. Lecture notes in statistics (Vol. 115, pp. 160–176). Berlin: Springer.

Davis, R. A., & Dunsmuir, W. T. M. (1996). Maximum likelihood estimation for MA(1) processes with a root on or near the unit circle. *Econometric Theory, 12*, 1–29.

Davis, R. A., & Mikosch, T. V. (2009). Probabilistic properties of stochastic volatility models. In T. G. Andersen, R. A. Davis, J.-P. Kreiss, & T. V. Mikosch (Eds.), *Handbook of financial time series* (pp. 255–268). Berlin: Springer.

de Gooijer, J. G., Abraham, B., Gould, A., & Robinson, L. (1985). Methods of determining the order of an autoregressive-moving-average process: A survey. *International Statistical Review, 53*, 301–329.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B, 39*, 1–38.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of American Statistical Association, 74*, 427–431.

Douc, R., Roueff, F., & Soulier, P. (2008). On the existence of some ARCH(∞) processes. *Stochastic Processes and Their Applications, 118*, 755–761.

Duong, Q. P. (1984). On the choice of the order of autoregressive models: a ranking and selection approach. *Journal of Time Series Analysis, 5*, 145–157.

Eberlein, E. (2009). Jump-type Lévy processes. In T. G. Andersen, R. A. Davis, J.-P. Kreiss, & T. V. Mikosch (Eds.), *Handbook of financial time series* (pp. 439–456). Berlin: Springer.

Eller, J. (1987). On functions of companion matrices. *Linear Algebra and Applications, 96*, 191–210.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation. *Econometrica, 50*, 987–1007.

Engle, R. F. (1995). *ARCH: Selected readings*. Advanced texts in econometrics. Oxford: Oxford University Press.

Engle, R. F., & Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Economic Review, 5*, 1–50.

Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation and testing. *Econometrica, 55*, 251–276.

Engle, R. F., & Granger, C. W. J. (1991). *Long-run economic relationships*. Advanced texts in econometrics. Oxford: Oxford University Press.

Francq, C., & Zakoian, J.-M. (2010). *GARCH models: Structure, statistical inference and financial applications*. New York: Wiley.

Fuller, W. A. (1976). *Introduction to statistical time series*. New York: Wiley.

Gourieroux, C. (1997). *ARCH models and financial applications*. New York: Springer.

Granger, C. W. J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics, 16*, 121–130.

Gray, H. L., Kelley, G. D., & McIntire, D. D. (1978). A new approach to ARMA modeling. *Communications in Statistics, B7*, 1–77.

Graybill, F. A. (1983). *Matrices with applications in statistics*. Belmont, CA: Wadsworth.

Grunwald, G. K., Hyndman, R. J., & Hamza, K. (1994). *Some Properties and Generalizations of Nonnegative Bayesian Time Series Models, Technical Report*. Statistics Dept., Melbourne University, Parkville, Australia.

Grunwald, G. K., Raftery, A. E., & Guttorp, P. (1993). Prediction rule for exponential family state space models. *Journal of the Royal Statistical Society B, 55*, 937–943.

Hannan, E. J. (1980). The estimation of the order of an ARMA process. *Annals of Applied Statistics, 8*, 1071–1081.

Hannan, E. J., & Deistler, M. (1988). *The statistical theory of linear systems*. New York: Wiley.

Hannan, E. J., & Rissanen, J. (1982). Recursive estimation of mixed autoregressive moving-average order. *Biometrika, 69*, 81–94.

Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*. Cambridge: Cambridge University Press.

Harvey, A. C., & Fernandes, C. (1989). Time Series models for count data of qualitative observations. *Journal of Business and Economic Statistics, 7*, 407–422.

Holt, C. C. (1957). Forecasting seasonals and trends by exponentially weighted moving averages. *ONR research memorandum* (Vol. 52). Pittsburgh, PA: Carnegie Institute of Technology.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*, 297–307.

Iacus, S.M. and Mercuri, L. (2015), Implementation of Lévy CARMA model in Yuima package, *Comput. Stat.*, 30, 1111–1141.

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, heteroscedasticity and serial independence of regression residuals. *Economics Letters, 6*, 255–259.

Jones, R. H. (1975). Fitting autoregressions, *Journal of American Statistical Association, 70*, 590–592.

Jones, R. H. (1978). Multivariate autoregression estimation using residuals. In D. F. Findley (Ed.), *Applied time series analysis* (pp. 139–162). New York: Academic.

Jones, R. H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics, 22*, 389–395.

Kendall, M. G., & Stuart, A. (1976). *The advanced theory of statistics* (Vol. 3). London: Griffin.

Kitagawa, G. (1987). Non-Gaussian state-space modeling of non-stationary time series. *Journal of the American Statistical Association, 82* (with discussion), 1032–1063.

Klebaner, F. (2005). *Introduction to stochastic calculus with applications*. London: Imperial College Press.

Klüppelberg, C., Lindner, A., & Maller, R. (2004). A continuous-time GARCH process driven by a Lévy process: stationarity and second-order behaviour. *Journal of Applied Probability, 41*, 601–622.

Kuk, A. Y. C., & Cheng, Y. W. (1994). *The Monte Carlo Newton-Raphson Algorithm, Technical Report S94-10*. Department of Statistics, U. New South Wales, Sydney, Australia.

Lehmann, E. L. (1983). *Theory of point estimation*. New York: Wiley.

Lehmann, E. L. (1986). *Testing statistical hypotheses* (2nd ed.). New York: Wiley.

Lindner, A. (2009). Stationarity, mixing, distributional properties and moments of GARCH(p,q)-processes. In T. G. Andersen, R. A. Davis, J.-P. Kreiss, & T. V. Mikosch (Eds.), *Handbook of financial time series* (pp. 233–254). Berlin: Springer.

Liu, J. & Brockwell, P. J. (1988). The general bilinear time series model. *Journal of Applied Probability, 25*, 553–564.

Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika, 65*, 297–303.

Lütkepohl, H. (1993). *Introduction to multiple time series analysis* (2nd ed.). Berlin: Springer.

Mage, D. T. (1982). An objective graphical method for testing normal distributional assumptions using probability plots. *American Statistician, 36*, 116–120.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1984). *The forecasting accuracy of major time series methods*. New York: Wiley.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1997). *Forecasting: Methods and applications*. New York: Wiley.

May, R. M. (1976). Simple mathematical models with very complicated dynamics. *Nature, 261*, 459–467.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman and Hall.

McLeod, A. I., & Li, W. K. (1983). Diagnostic checking ARMA time series models using squared-residual autocorrelations. *Journal of Time Series Analysis, 4*, 269–273.

Mendenhall, W., Wackerly, D. D., and Scheaffer, D. L. (1990). *Mathematical statistics with applications* (4th ed.). Belmont: Duxbury.

Merton, R. (1973). The theory of rational option pricing. *Bell Journal of Economics and Management Science, 4*, 141–183.

Mikosch, T. (1998), *Elementary stochastic calculus with finance in view*. Singapore: World Scientific.

Mood, A.M., Graybill, F.A. and Boes, D.C. (1974), Introduction to the Theory of Statistics, McGraw-Hill, New York.

Nelson, D. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica, 59*, 347–370.

Newton, H. J., & Parzen, E. (1984). Forecasting and time series model types of 111 economic time series. In S. Makridakis, et al. (Eds.), *The forecasting accuracy of major time series methods*. New York: Wiley.

Nicholls, D. F., & Quinn, B. G. (1982). *Random coefficient autoregressive models: An introduction*. Springer lecture notes in statistics (Vol. 11), Springer-Verlag, Berlin, Heidelberg, New York.

Oksendal, B. (2013). *Stochastic differential equations: An introduction with applications* (6th ed.). New York: Springer.

Pantula, S. (1991). Asymptotic distributions of unit-root tests when the process is nearly stationary. *Journal of Business and Economic Statistics, 9*, 63–71.

Parzen, E. (1982), ARARMA models for time series analysis and forecasting. *Journal of Forecasting, 1*, 67–82.

Pole, A., West, M., & Harrison, J. (1994). *Applied Bayesian forecasting and time series analysis*. New York: Chapman and Hall.

Priestley, M. B. (1988). *Non-linear and non-stationary time series analysis*. London: Academic.

Protter, P. E. (2010). *Stochastic integration and differential equations* (2nd ed.). New York: Springer.

Rosenblatt, M. (1985). *Stationary sequences and random fields*. Boston: Birkhäuser.

Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive moving-average models with unknown order. *Biometrika, 71*, 599–607.

Sakai, H., & Tokumaru, H. (1980). Autocorrelations of a certain chaos. In *IEEE Transactions on Acoustics, Speech and Signal Processing* (Vol. 28, pp. 588–590).

Samuelson, P. A. (1965). Rational theory of warrant pricing. *Industrial Management Review, 6*, 13–31.

Sato, K. (1999). *Lévy processes and infinitely divisible distributions* . Cambridge: Cambridge University Press.

Schoutens, W. (2003). *Lévy processes in finance*. New York: Wiley.

Schwert, G. W. (1987). Effects of model specification on tests for unit roots in macroeconomic data. *Journal of Monetary Economics, 20*, 73–103.

Shapiro, S. S., & Francia, R. S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association, 67*, 215–216.

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. In D. R. Cox, D. V. Hinkley, & O. E. Barndorff-Nielsen (Eds.), *Time series models in econometrics, finance and other fields* (pp. 1–67). London: Chapman and Hall.

Shephard, N., & Andersen, T. G. (2009). Stochastic volatility: Origins and overview. In T. G. Andersen, R. A. Davis, J.-P. Kreiss, & T. V. Mikosch (Eds.), *Handbook of financial time series* (pp. 233–254). Berlin, Heidelberg: Springer.

Shibata, R. (1976), Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika, 63*, 117–126.

Shibata, R. (1980), Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics, 8*, 147–164.

Silvey, S. D. (1975). *Statistical inference*. New York: Halsted.

Smith, J. Q. (1979). A generalization of the Bayesian steady forecasting model. *Journal of the Royal Statistical Society B, 41*, 375–387.

Sorenson, H. W., & Alspach, D. L. (1971). Recursive Bayesian estimation using Gaussian sums. *Automatica, 7*, 465–479.

Subba-Rao, T., & Gabr, M. M. (1984). *An introduction to bispectral analysis and bilinear time series models*. Springer lecture notes in statistics (Vol. 24), Springer-Verlag, Berlin, Heidelberg, New York.

Tam, W. K., & Reinsel, G. C. (1997). Tests for seasonal moving-average unit root in ARIMA models. *Journal of the American Statistical Association, 92*, 725–738.

Tanaka, K. (1990). Testing for a moving-average unit root. *Econometric Theory, 9*, 433–444.

Taylor, S. J. (1982). Financial returns modelled by the product of two stochastic processes-a study of the daily sugar prices 1961–75. *Time Series Analysis: Theory and Practice, 1*, 203–226.

Taylor, S. J (1986). *Modelling financial time series*. New York: Wiley.

Tong, H. (1990). *Non-linear time series: A dynamical systems approach*. Oxford: Oxford University Press.

Venables, W. N., Ripley, B. D. (2003). *Modern applied statistics with S* (4th ed.). New York: Springer.

Weigt, G. (2015), ITSM-R Reference Manual. The manual can be downloaded from http://eigenmath.sourceforge.net/itsmr-refman.pdf.

Weiss, A. A. (1986). Asymptotic theory for ARCH models: Estimation and testing. *Econometric Theory, 2*, 107–131.

West, M., & Harrison, P. J. (1989). *Bayesian forecasting and dynamic models*. New York: Springer.

Whittle, P. (1963). On the fitting of multivariate autoregressions and the approximate canonical factorization of a spectral density matrix. *Biometrika, 40*, 129–134.

Wichern, D., & Jones, R. H. (1978). Assessing the impact of market disturbances using intervention analysis. *Management Science, 24*, 320–337.

Wu, C. F. J. (1983). On the convergence of the EM algorithm. *Annals of Statistics, 11*, 95–103.

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika, 75*, 621–629.

Index

P.J. Brockwell, R.A. Davis, *Introduction to Time Series and Forecasting*,
Springer Texts in Statistics, DOI 10.1007/978-3-319-29854-2